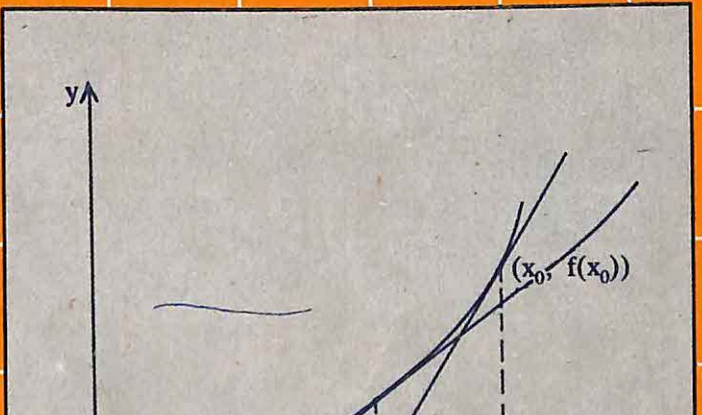
MATHEMATICS

A Textbook for Class XII

PART III



राष्ट्रीय शैक्षिक अनुसंधान और प्रशिक्षण परिषद्
NATIONAL COUNCIL OF EDUCATIONAL RESEARCH AND TRAINING

9934
165.24

MATHEMATICS

A Textbook for Class XII

Part III

D. D. JOSHI U. B. TEWARI
K. V. RAO V. K. KANNAN
M. S. RANGACHARI V. G. TIKEKAR
S. IZHAR HUSAIN

Received
Cont. n s not verified
...5.22......Dt..25/4/97...
SCERT, West Bengal

राष्ट्रीय शैक्षिक अनुसंधान और प्रशिक्षण परिषद्
NATIONAL COUNCIL OF EDUCATIONAL RESEARCH AND TRAINING

First Edition
April 1989 Vaisakha 1911
Fifth Reprint Edition
May 1993 . Vaisakha 1915
First Revised Edition
June 1994 Jyaistha 1916

PD 160T YPS

9934
16.5.2K

© *National Council of Educational Research and Training, 1994*

ALL RIGHTS RESERVED

☐ No part of this publication may be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without the prior permission of the publisher.

☐ This book is sold subject to the condition that it shall not, by way of trade, be lent, re-sold, hired out or otherwise disposed of without the publisher's consent, in any form of binding or cover other than that in which it is published.

☐ The correct price of this publication is the price printed on this page. Any revised price indicated by a rubber stamp or by a sticker or by any other means is incorrect and should be unacceptable.

[illegible] LIBRARY

No ....

Publication Team

C. N. Rao *Head, Publication Department*

Prabhakar Dwivedi *Chief Editor*
J. P. Sharma *Editor*
Y. P. Sharma *Assistant Editor*
R. N. Bhardwaj *Editorial Assistant*

U. Prabhakar Rao *Chief Production Officer*
Suresh Chand *Production Officer*
S. M. Muslim *Assistant Production Officer*
D. K. Shende *Senior Artist*
Parkash Tahilyani *Production Assistant*

Cover Deepesh Sharma

OFFICES OF THE PUBLICATION DEPARTMENT, NCERT

NCERT Campus
Sri Aurobindo Marg
NEW DELHI 110016

CWC Campus
Chitlapakkam, Chromepet
MADRAS 600064

Navjivan Trust Building
P.O. Navjivan
AHMEDABAD 380014

CWC Campus
32, B.T. Road, Sukchar
24 PARGANAS 743179

Rs 17.50

Published at the Publication Department by the Secretary, National Council of Educational Research and Training, Sri Aurobindo Marg, New Delhi 110 016, lasertypeset at Computer Data System, 270, Satya Niketan, Ring Road, New Delhi 110 021 and printed at The Coronation Litho Works, 38-39, Thiruthangal Road, Sivakasi 626 123

Contents

11. **Probability** 551
  11.1 Introduction 551
  11.2 Random Experiments 551
  11.3 Sample Space 552
  11.4 Events 554
  11.5 Algebra of Events 557
  11.6 Probability of an Event 560
  11.7 Theorems on Probability 562
  11.8 Use of Permutations and Combinations in Calculation of Probabilities 566
  11.9 General Definition of a Random Experiment 569
  11.10 Conditional Probability 573
  11.11 Independent Events 577
  11.12 Independent Experiments 580
  11.13 The Law of Total Probability and Bayes' Formula 586
  11.14 Random Variables 590
  11.15 Probability Distribution of a Random Variable 592
  11.16 Binomial Distribution 595

12. **Correlation and Regression** 603
  12.1 Bivariate Frequency Distributions 603
  12.2 Marginal Distributions 607
  12.3 Conditional Distributions 608
  12.4 Relationship between Variables 609
  12.5 Covariance 611
  12.6 Calculation of Covariance 612
  12.7 Meaning of Covariance 614
  12.8 Coefficient of Correlation 620
  12.9 Calculation of the Correlation Coefficient 622
  12.10 Interpretation of the Correlation Coefficient 624
  12.11 Introduction 627
  12.12 Least Squares Lines of Regression 628
  12.13 Calculation of Regression Coefficients 631
  12.14 Relationship with Correlation Analysis 633

Received
Contents not verified
..................Dt..................
SCERT, West Bengal

13. Computing **638**
13.1 Introduction 638
13.2 Block Diagram of a Computer 638
13.3 Memory 639
13.4 Number Systems 640
13.5 Computer Arithmetic 647
13.6 Algorithm 657
13.7 Variable Name or Indentifier 661
13.8 Assignment Operation 662
13.9 Assignment Symbol to Keep Count 663
13.10 A Pseudo Language 665
13.11 Formal Explanation of Constructs 668
13.12 Stepwise Refinement of Algorithm 673
13.13 How an Algorithm is Presented ? 678
13.14 Some More Examples 680

14. Numerical Methods **684**
14.1 Introduction 684
14.2 Successive Bisection Method 685
14.3 The Method of False Position 692
14.4 Newton-Raphson Method 698
14.5 Solution of a System of Equations 701
14.6 Approximations of Some Functions 710
14.7 Numerical Integration 712

ANSWERS **721**

Chapters 1 – 10 have been covered in Parts I and II of the book.

CHAPTER 11

Probability

A. Random Experiments and Probability

11.1 Introduction

The theory of probability owes its origins to the study of games of chance or gambling. Suppose we are permitted to toss a coin on payment of Rs 2 and are offered Rs 3 as a prize if the throw of the coin results in head. If the coin shows tail we do not get anything. Should we play such a betting game or not ? Or, suppose two dice are thrown in turn by two players A and B. If the total of the numbers on the two dice is more than 6, A pays Re 1 to B, if it is 6 or less, B pays Re 1 to A. Is this betting game more favourable to A than to B ? These are the types of questions which mathematicians tried to answer which led to the development of the modern theory of probability.

The distinctive feature of games of chance is that we are faced with situations where under the given conditions more than one result is possible. When we throw two dice the total of the numbers appearing on them can be any number from 2 to 12. Although we know what the possible results are, we are not sure which one of these results will actually appear. Likewise, when we ask the question: "Is it going to rain tomorrow?", we are referring to a phenomenon which may or may not occur.

Probability theory is designed to deal with uncertainties regarding the happening of given phenomena. The word "probable" itself indicates such a situation. Its dictionary meaning is "likely though not certain to occur". Thus when a coin is thrown a head is likely to occur but may not occur. When a die is thrown it may or may not show the number 3. In the same way, when we talk of the list of probables for the national cricket or hockey team we are referring to a list of persons who are likely to play for the country but are not certain to be included in the team.

The aim of probability theory is to provide a mathematical model to study uncertain situations in the same manner as geometry provides a mathematical theory for dealing with practical problems concerning areas, volumes and space.

11.2 Random Experiments

The basic idea to begin with is that of a *random experiment*. The expression "random experiment" describes a general mathematical concept which covers all the particular cases which arise when one deals with uncertain situations.

The word "random" used above means "haphazard, occurring in an unplanned manner, etc.". Instead of using the expression "random experiment ", we could as well have used the expressions "chance experiment" or "probability experiment" . We continue to use the term random experiment for historical reasons only, as this expression has been established through a long period of use. The word random will occur in other contexts also and only be an indicator of the fact that uncertain situations are involved.

Examples of random experiments are: tossing a coin, throwing a die, selecting a card from a pack of playing cards, selecting a family out of a given group of families, etc. In all these cases there are a number of possible results which can occur but there is an uncertainty as to which one of them will actually occur.

For the present we will say that a random experiment has been *defined* when all possible outcomes have been listed. What this means is that for the purpose of developing a mathematical theory we can ignore the actual physical description of the random experiment once we have described the set of its outcomes. All further ideas can be developed from this set of outcomes alone. The definition of a random experiment will be given later.

11.3 Sample Space

We give the name *sample space* to the set of all possible outcomes of a random experiment. Actually, we should use the expression " sample space associated with a random experiment", since every sample space refers to a given random experiment. But we shall mostly use the shortened expression "sample space"only, the experiment to which it refers to being clear from the context.

When we toss a coin there are two possible outcomes only: a head or a tail. The sample space of this experiment (tossing a coin) consists of two elements and may be written as the set $\{H, T\}$. Similarly, when we throw a die it can result in any of the six numbers 1, 2, 3, 4, 5, 6. The sample space of this experiment consists, therefore, of six elements and may be described by the set $\{1, 2, 3, 4, 5, 6\}$.

Suppose we toss two coins. How shall we describe the sample space of this experiment ? We may think of the sample space as consisting of three outcomes only :"no heads", "one head only", "two heads", or, as "no tails", "one tail only", "two tails". If we can somehow distinguish one coin from the other, by tossing them one after the other, for example, the sample space can be seen to consist of four outcomes:

"head on first coin, head on second coin",

"head on first coin, tail on second coin",

"tail on first coin, head on second coin" and

"tail on first coin, tail on second coin".

This sample space of four outcomes may be described by the set

$$\{(H, H), (H, T), (T, H), (T, T)\},$$

where (H, T) stands for the *ordered pair* : first H, then T.

Even if we do not distinguish one coin from the other we describe the sample space of the experiment of tossing two coins by the set $\{(H, H), (H, T), (T, H), (T, T)\}$. The reason for doing so may become clear by considering the example of the experiment of tossing a coin and a die together. Here the sample space consists of 12 outcomes consisting of a head or tail occuring with any of the numbers from 1 to 6. This sample space can be described by the set

$$\{(H, 1), (H, 2), (H, 3), (H, 4), (H, 5), (H, 6),$$
$$(T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\}.$$

More generally, suppose we have a random experiment with m outcomes $x_1, x_2, ..., x_m$ and another with n outcomes $y_1, y_2, \ldots, y_n$; then the sample space of the experiment which consists of carrying out the two experiments together has mn outcomes and can be described by the set

$$\begin{aligned}\{&(x_1, y_1), (x_1, y_2), ..., (x_1, y_n),\\ &(x_2, y_1), (x_2, y_2), \ldots, (x_2, y_n),\\ &..., ..., ..., ..., ..., ..., ..., ...,\\ &(x_m, y_1), (x_m, y_2), \ldots, (x_m, y_n)\}\end{aligned}$$

or, in short, by the set

$$\{(x_i, y_j) : i = 1, \ldots, m; j = 1, \ldots, n\}.$$

We shall denote the sample space of a random experiment by S and the outcomes of the experiment, that is, the elements of set S, by $\omega_1, \omega_2, \ldots$.

Example 11.1

From a group of 3 boys and 2 girls we select two children. What would be the sample space of this experiment ?

Solution

We may say that there are only three possible outcomes : two boys are selected, one boy and one girl is selected, 2 girls are selected. We could also say that since two children out of five can be selected in $C(5, 2) = 10$ ways, the number of possible outcomes is 10. If we designate the boys as B_1, B_2, B_3 and the girls as G_1, G_2, the possible outcomes can be described easily by the following diagram shown in Fig. 11.1.

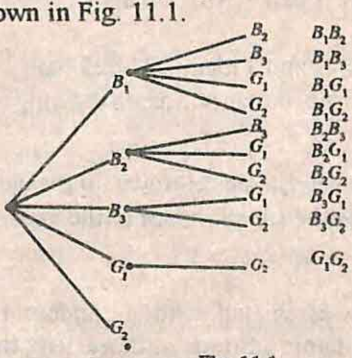


Fig. 11.1

Thus the sample space of the experiment is

$$\{B_1B_2, B_1B_3, B_1G_1, B_1G_2, B_2B_3,$$
$$B_2G_1, B_2G_2, B_3G_1, B_3G_2, G_1G_2\}.$$

Thus, the sample space in this experiment should be taken as a set consisting of 10 elements.

Example 11.2

A coin is tossed. If it shows head, we draw a ball from a bag consisting of 3 red and 4 black balls; if it shows tail, we throw a die. What is the sample space of this experiment ?

Solution

We denote the red balls by r_1, r_2, r_3 and the black balls by b_1, b_2, b_3, b_4. Then the sample space is the set

$$\{Hr_1, Hr_2, Hr_3, Hb_1, Hb_2, Hb_3, Hb_4,\\ T1, T2, T3, T4, T5, T6\}$$

and consist of 13 elements.

EXERCISE 11.1

1. If three coins are tossed simultaneously, what is the resulting sample space ?
2. A coin is tossed twice. If the second throw results in a tail, a die is thrown. Describe the sample space.
3. A coin is tossed twice. If the second throw results in a head, a die is thrown; otherwise a coin is tossed. What is the sample space ?
4. A bag contains 4 red balls. What is the sample space if the experiment consists of choosing (a) 1 ball ? (b) 2 balls ? (c) 3 balls ? (d) 4 balls ?
5. A bag contains 4 red balls and 3 black balls. What is the sample space if the experiment consists of drawing from the bag:

   (a) 1 ball ? (b) 2 balls ?
6. A bag contains 4 identical red balls and 3 identical black balls. The experiment consists of drawing one ball, then putting it into the bag and again drawing a ball. What are the possible outcomes of the experiment ?
7. A coin is tossed. If it results in a head, a die is thrown. If the die shows up an even number, the die is again thrown. What is the sample space of the experiment ?

11.4 Events

The elements of the sample space associated with a random experiment are called the *elementary events* of that experiment. Some authors use the term *simple event* or *indecomposable event*. Thus, the experiment of tossing a coin has two elementary events which may be denoted by H and T. Similarly, the six elementary events of the experiment of throwing a die may be denoted by 1, 2, 3, 4, 5 and 6.

However, the notion of an event is more general. When we throw a die and ask the question : Is 2 the number shown ?, the answer would be Yes or No. Similarly, every question asking if the die has come up with one of the numbers from 1 to 6 (i.e. a number which is an elementary event of the experiment) has Yes or No as the answer. There are many other questions which have Yes or No as the answer; for example, the question : is the number shown an even number, is it less than 5, is it greater than 2, and so on.

We say that any question of this type defines an *event* on the sample space of the experiment. In the experiment which consists of tossing two coins the questions: do both coins show heads, do both coins show the same result, is the number of heads same as the number of tails, etc., all have Yes or No as an answer, and each one of these questions defines an event on the sample space of the experiment of tossing two coins. If the answer to the question is Yes, we say that the event defined by the question has occurred; if the answer is No, we say that the event has not occurred.

In the experiment of throwing a die consider the event defined by the question: Is the number shown an even number ? If the throw of the die result in 2, we will say that the event has occurred, but if the die shows the number 5, we will say that the event has not occurred. Out of the six possible outcomes of the experiment, three (the numbers 2, 4, 6) indicate that the event has occurred, while the other three (the numbers 1, 3, 5) indicate that the event has not occurred. Thus instead of defining the event through a question, we may as well describe it by the statement: " The number shown at the throw of the die is an even number", and denote it by the subset $\{2, 4, 6\}$ of the sample space with the rule that if the result of the experiment belongs to this subset then the event is said to have occurred. And, if the result of the experiment does not belong to this subset, i.e. belongs to the subset $\{1, 3, 5\}$, we say that the event has not occurred.

Similarly, in the case of the experiment of tossing two coins, we may talk of the event: "both coins show the same result" which would be represented by the subset $\{(H, H), (T, T)\}$ of the sample space $\{(H, H), (H, T), (T, H), (T, T)\}$. For the same experiment the event: "the number of heads is the same as the number of tails" will be represented by the subset $\{H, T), (T, H)\}$.

In general, given the sample space S of a random experiment, *every* event associated with the experiment can be represented by a subset E of S. Conversely, *every* subset E of S represents an event associated with the random experiment. The event represented by the subset E of S is said to have occurred if the outcome ω of the experiment is such that $\omega \in E$. If the outcome ω is such that $\omega \notin E$, we say that the event represented by the subset E has not occurred. We may thus talk of the event E instead of the event represented by the subset E and use the same symbol E for the event as well as the subset of the sample space representing the event.

Suppose the throw of a die results in the number 4. Then, on the basis of this outcome a number of events can be said to have occurred. For example, we can say that the following events have occurred:

(i) The number is even, represented by the subset $\{2, 4, 6\}$,
(ii) The number is more than 2, represented by the subset $\{3, 4, 5, 6\}$, and
(iii) The number is less than 5, represented by the subset $\{1, 2, 3, 4\}$.

On the basis of the same outcome, we can also say that a number of events have not occurred. For example, the following events have not occurred:

(i) The number is odd, represented by the subset $\{1, 3, 5)$,
(ii) The number is less than 3, represented by the subset $\{1, 2\}$, and
(iii) The number is more than 5, represented by the subset $\{6\}$.

The set S is itself a subset of S and hence can be thought of as representing an event associated with the experiment which has S as its sample space. Since every outcome of the experiment belongs to S, we will conclude that the event represented by S has occurred whatever be the outcome of the experiment. In other words, the event represented by S always occurs when the experiment is carried out. For this reason the event represented by S is called the *sure event*.

In the same manner the empty set ϕ, which is also a subset of S can be taken to represent an event. However, since no outcome of the experiment can belong to ϕ, the event represented by ϕ is such that it cannot occur at all when the experiment is performed. The event represented by ϕ is, therefore, called the *impossible event*.

Every other subset of S which is different from ϕ and S represents an event which may or may not occur when the experiment is carried out.

Example 11.3

In the sample space of Example 11.1, the event: "both selected children are boys" is represented by the set $\{B_1B_2, B_1B_3, B_2B_3\}$ and the event: "the selected group consists of one boy and one girl" is represented by the set

$$\{B_1G_1, B_1G_2, B_2G_1, B_2G_2, B_3G_1, B_3G_2\}.$$

Example 11.4

In the sample space of Example 11.2, the event: "the throw of the coin resulted in head" is represented by the set

$$\{Hr_1, Hr_2, Hr_3, Hb_1, Hb_2, Hb_3, Hb_4\}$$

and the event: "the throw of a die resulted in an even number" is represented by the set $\{T2, T4, T6\}$.

EXERCISE 11.2

1. In the sample space of Example 11.2, what are the sets representing the events
   (a) A red ball is drawn ?
   (b) The throw of the coin resulted in tail ?
2. In the sample space of Example 11.1, what is the set representing the event: "at least one girl is selected" ?

3. A coin and a die are tossed. Describe the following events:

   (i) A = getting a head and an even number
   (ii) B = getting a prime number
   (iii) C = getting a tail and an odd number
   (iv) D = getting a head or a tail.

4. A coin is tossed. If it results in a head a coin is tossed, otherwise a die is thrown. Describe the following events:

   (i) A = getting at least one head
   (ii) B = getting an even number
   (iii) C = getting a tail
   (iv) D = getting a tail and an odd number

11.5 Algebra of Events

Given some events associated with a random experiment, we can define new events in terms of them. We will now discuss some standard methods of doing so.

Consider two events: "the number is even" and "the number is more than 3" associated with the random experiment of throwing a die. The sets E and F representing these events are $E = \{2, 4, 6\}$ and $F = \{4, 5, 6\}$. We now define a new event "E or F" which occurs when E or F or both occur. What is the subset which represents this new event ? It is clear that the numbers 2, 4, 6 belong to this subset and so also the numbers 4, 5, 6. At the same time no other outcome, i.e. 1 or 3, can belong to this subset. Thus, the event "E or F" will be represented by the subset $E \cup F = \{2, 4, 5, 6\}$.

We can also define a new event "E and F" which occurs only when E and F both occur. If the outcome is 2, the event E occurs but F does not. Hence, it is clear that the event "E and F" can occur only when the outcome belongs to E and F both. Thus, the event "E and F" will be represented by the set $E \cap F = \{4, 6\}$.

Thus, with any two events represented by the subsets E and F of a sample space S we can associate two new events defined by the conditions "either E or F or both occur" and "both E and F occur". These events will be called the events "E or F" and "E and F" and will be represented by the subsets $E \cup F$ and $E \cap F$ respectively.

In the example considered above the events "the number is even" and "the number is more than 3" were such that on the basis of some outcomes both could have occurred. For example, if the outcome were 6 (or 4) we would say that both the events have occurred. However, the events "the number is less than 3" and "the number is more than 5" are such that both of them cannot be said to have occurred together whatever be the outcome of the experiment. For, the first event occurs when the outcome is 1 or 2, and the second occurs when the outcome is 6. Such events, where the occurrence of one precludes the occurrence of the other, are called *mutually exclusive events*. It is clear that if two events associated with the same experiment are mutually exclusive the subsets of the sample space representing the two events are disjoint. Conversely, if the subsets are disjoint they represent mutually exclusive events.

Consider next, the events "the number is even" and "the number is odd" associated with the random experiment of throwing a die. The two events are obviously mutually exclusive but we can say something more. Here at least one of the events has to occur. In other words, if the first event does not occur the second must occur, and the non-occurrence of the second means the first event must have occurred. Given an event E, the event which occurs when, and only when, E does not occur is called the event "not-E ". If the event E is represented by the subset E of the sample space **S**, the event "not-E " will be represented by the subset consisting of all the elements of **S** which do not belong to E. That is to say, the event "not-E " will be represented by E^c the *complement* (in **S**) of the set E. For this reason the event "not-E" is also called the *complementary* event of E. Some authors also call it the *negation* of E.

The events E and not-E are such that only one of them can occur, and at least one of them must occur. Such a situation can arise even when we have more than two events. For example, let the experiment be that of drawing a card from a pack of fifty-two playing cards. The 52 cards of the pack are divided into four types of cards, each type consisting of 13 cards. These four types are given the names: spades, hearts, diamonds and clubs. Spades and clubs cards are black in colour while the other two types are red. Associated with this experiment we may, therefore, define the following four events: "card drawn is spades", "card drawn is hearts", "card drawn is diamonds", and "card drawn is clubs". Now, one of these events must occur since the card drawn is necessarily one of the four types. At the same time, if any one of these events occurs the others cannot occur. We call such a collection of events as forming a *mutually exclusive and exhaustive system of events*. If E_1, E_2, ..., E_k are the subsets of a sample space **S** representing a system of mutually exclusive and exhaustive system of events, then we have

(i) $E_i \cap E_j = \phi$ for $i \neq j$, and
(ii) $E_1 \cup E_2 \cup \ldots \cup E_k = \mathbf{S}$.

In set theory we also have the *relation* "A is a subset of B" which we write as $A \subset B$. Now, suppose A and B are subsets of a sample space **S** and we have $A \subset B$. The subsets A and B represent two events associated with the random experiment with sample space **S**. The event A will be said to have occurred when the outcome ω of the experiment is such that $\omega \in A$. But in that case we also have $\omega \in B$ as $A \subset B$ and B will also be said to have occurred. So, events A and B are such that if we know that the event A has occurred we can also say that the event B has occurred. In the language of probability theory we say that "event B is *implied* by the event A". This statement is just a standard way of saying that if the event A occurs then event B must also occur.

We have seen that the events and the relationships among them are best described by using the language of set theory. We give below the equivalent forms of probability statements and the corresponding statements in the notation of set theory.

Probability theory	*Set notation*
Sample space	$\mathbf{S}$
Outcome of the random experiment	ω
Event A	$A \subset \mathbf{S}$
Event A has occurred	$\omega \in A$
Event A has not occurred	$\omega \notin A$
Event "A or B"	$A \cup B$
Event "A and B"	$A \cap B$
Event "not-A"	A^c
Event A implies event B	$A \subset B$
Events A and B are mutually exclusive	$A \cap B = \phi$
Events $A_1, \ldots, A_m$ are mutually exclusive	$A_i \cap A_j = \phi$ for $i \neq j$
and exhaustive	and $\bigcup_{i=1}^{m} A_i = \mathbf{S}$

Example 11.5

If events E and F are represented by the subsets E and F of the sample space S of a random experiment, the events

(i) only E occurs,

(ii) only F occurs,

(iii) none of them occurs,

(iv) at least one of them occurs

are represented by the sets

$$E \cap F^c,\ E^c \cap F,\ E^c \cap F^c \text{ and } E \cup F$$

respectively. Note that events (iii) and (iv) are complementary events which is also seen from the fact that $(E \cup F)^c = E^c \cap F^c$.

EXERCISE 11.3

1. Two dice are thrown. The events A, B, C, D, E, F are as follows:

A = getting an even number on the first die.

B = getting an odd number on the first die.

C = getting the sum of the numbers on the dice ≤ 5

D = getting the sum of the numbers on the dice greater than 5 but less than 10.

E = getting the sum of the numbers on the dice ≥ 10.

F = getting an odd number on one of the dice.

(a) Describe the following events:

(i) A^c (ii) B^c (iii) E^c (iv) A or B
(v) A and B (vi) B or C (vii) B and C
(viii) A and E (ix) A or F (x) A and F.

(b) State True or False:

(i) A and B are mutually exclusive.
(ii) A and B are mutually exclusive and exhaustive events.
(iii) $A = B^c$
(iv) A and C are mutually exclusive.
(v) C and D are mutually exclusive.
(vi) C and D are mutually exclusive and exhaustive.
(vii) $D^c = C$
(viii) C, D, E are mutually exclusive and exhaustive events.
(ix) A^c and B^c are mutually exclusive and exhaustive.
(x) A, B, F are mutually exclusive and exhaustive events.

2. Three coins are tossed.

(i) Describe two events A and B which are mutually exclusive.
(ii) Describe three events A, B and C which are mutually exclusive and exhaustive.
(iii) Describe two events A and B which are *not* mutually exclusive.
(iv) Describe three events A, B, C which are *not* mutually exclusive.
(v) Describe two events which are mutually exclusive but *not* exhaustive.
(vi) Describe three events, A, B and C which are mutually exclusive, but *not* exhaustive.

3. If $E_1, E_2, \ldots, E_n$ are a set of mutually exclusive and exhaustive set of events, prove that the non-occurrence of E_1, E_2 or E_3 means that at least one of the remaining events $E_4, E_5, \ldots, E_n$ must have occurred.
(Hint: $E_1 \cup E_2 \cup E_3$ and $E_4 \cup E_5 \cup \ldots \cup E_n$ are complementary events.)

11.6 Probability of an Event

Suppose we have a bag full of small balls which are similar in shape and size but some are black in colour and the other are red. If the balls are thoroughly mixed and then one ball is drawn it will be either black or red. If you were asked to guess whether the ball drawn is red or black then, in the absence of any additional informaiton, you would say that the ball drawn is as likely to be red as black. But if you were told that there are 100 balls in the bag of which only one is red then you would say that the ball drawn is most likely to be black and that it is hardly possible that it would be red. If in fact the ball drawn happens to be red you would be surprised at something unexpected happening. If, on the other hand, the bag were to consist of 30 red and 70 black balls you would not be surprised to see a red ball turning up at the draw though you would still feel that the ball drawn is more likely to be black.

In the same way, if we have two events E and F associated with a random experiment, we can talk of the event E being more or less likely to happen than the event F when the experiment is carried out. We say that the event which is more likely to occur has a higher *probability* than an event which is less likely to occur. For example, in the experiment of drawing a card from a pack of 52 playing cards, the event "the card drawn is black" can be said to have a higher probability than the event "the card drawn is an Ace".

To every event associated with a random experiment we try to attach a numerical value called its *probability* in such a manner that for any two events the event which is more likely to happen has a higher value for the probability. Since every event is more likely to happen than the impossible event, and less likely to happen than the sure event, the impossible event should have the smallest probability and the sure event should have the largest probability. By convention, the numerical value of the probability of the impossible event is taken as zero and of the sure event as 1, the probabilities of all other events having values between 0 and 1. We shall denote the probability of the event E by the symbol $P\,(E)$.

Suppose a random experiment has n outcomes so that the sample space S has n elements. Let E be the subset of S representing an event and let us use the symbol E to denote the event also. Suppose the set E consists of m elementary events. Then, one way to attach a probability to every event is to define $P\,(E)$ as $P\,(E) = \frac{m}{n}$.

If $P\,(E)$ is defined as above, we see immediately that $P\,(\phi) = 0$, $P\,(\mathrm{S}) = 1$, and $0 \leq P\,(E) \leq 1$. Thus, this method of defining the probability of an event satisfies the requirement mentioned earlier that the value of the probability of any event should lie between 0 and 1 with the impossible event having probability 0 and the sure event having probability 1. If the number of elementary events in a set F is more than that in a set E, we will have $P\,(F) > P\,(E)$. Thus, our definition also assigns a higher probability to an event which we consider more likely to happen. For example, if a box contains 30 red and 70 black balls, we have

$$P\,(\text{ball drawn is red}) = \frac{30}{100}$$

and
$$P\,(\text{ball drawn is black}) = \frac{70}{100}$$

which agrees with our feeling that the ball drawn is more likely to be black than red.

By the above definition each elementary event has the same probability $\frac{1}{n}$. Thus, if we use the above method for defining the probabilities of events we are implicitly *assuming* that we regard all the elementary events as *equally likely* to occur when the experiment is performed.

The above definition of probability of an event is usually expressed in slightly different language. We call those elementary events which belong to the subset E of the sample space representing an event as *favourable* to occurrence of the given event. We may then say that an event occurs if and only if, the outcome of the random experiment is one of the elementary events favourable to the occurrence of the event. In view of this, the earlier definition of the probability $P(E)$ of an event E can be written as

$$P(E) = \frac{\text{No. of elementary events favourable to } E}{\text{Total No. of equally likely elementary events}}$$

We emphasise that this definition of $P(E)$ is the same as that given earlier, only the language used is different.

11.7 Theorems on Probability

Theorem 11.1

If E and F are two mutually exclusive events of a random experiment, the probability of occurrence of the event "E or F" is the sum of the probabilities of the events E and F or,

$$P(E \text{ or } F) = P(E) + P(F)$$

if the events E and F are mutually exclusive.

Proof

Let r, s be the number of elementary events respectively in the sets E and F representing the mutually exclusive events E and F. Let n be the total number of elementary events in the sample space S of which E and F are subsets. Since, E and F are mutually exclusive events, the sets E and F representing them are disjoint so that the number of elementary events in the set $E \cup F$ is $r + s$. Hence,

$$P(E \text{ or } F) = P(E \cup F) = \frac{r+s}{n}$$

$$= \frac{r}{n} + \frac{s}{n} = P(E) + P(F).$$

The above theorem is usually called the *addition theorem* or *addition rule* for probabilities of mutually exclusive events. In the language of subsets of the sample space S, the theorem is usually written as follows:

"If $E \cap F = \phi$, then $P(E \cup F) = P(E) + P(F)$".

Corollary

If $E_1, E_2, \ldots, E_k$ are mutually exclusive events, then

$$P(E_1 \cup E_2 \cup \ldots \cup E_k) = \sum_{i=1}^{k} P(E_i).$$

The following is a more general form of Theorem 11.1 which also covers events which are not mutually exclusive.

Theorem 11.2

If E and F are two events associated with a random experiment, then

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F).$$

Proof

Consider the Venn diagram below of subsets E and F of **S** which represent the two events, where the shaded portion represents the set $E \cap F$ (Fig. 11.2).

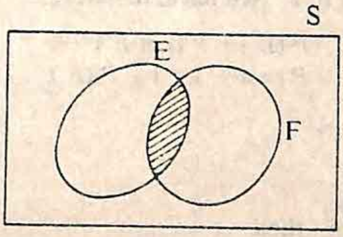


Fig. 11.2

Let r, s and t be the number of elementary events in the sets E, F and $E \cap F$ respectively and n, the total number of elementary events. The number of elementary events in the set $E \cup F$ representing the event "E or F" is then equal to $r + s - t$. Therefore,

$$P(E \text{ or } F) = \frac{r+s-t}{n} = \frac{r}{n} + \frac{s}{n} - \frac{t}{n}$$

$$= P(E) + P(F) - P(E \text{ and } F)$$

which proves the theorem.

In the language of the theory of sets we can write this result as

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

or as,

$$P(E) + P(F) = P(E \cup F) + P(E \cap F).$$

If the events E, F are mutually exclusive, they cannot occur together so that the event "E and F" is the impossible event with probability zero. Thus, for mutually exclusive events we get as before

$$P(E \text{ or } F) = P(E) + P(F).$$

Because of this we say Theorem 11.2 is a more general form of Theorem 11.1, or that, Theorem 11.1 is a special case of Theorem 11.2.

Theorem 11.3

For every event E associated with a random experiment we have

$$P(\text{not-}E) = 1 - P(E).$$

Proof

Events E and "not-E" are mutually exclusive. Hence

$$P(E \text{ or not-}E) = P(E) + P(\text{not-}E).$$

But one of the two events E and "not-E" must occur so that the event "E or not-E" is the sure event with probability equal to 1. We have, therefore,

$$P(E) + P(\text{not-}E) = 1$$

i.e., $$P(\text{not-}E) = 1 - P(E)$$

which proves the theorem.

Theorem 11.4

If the event E implies the event F, then

$$P(E) \le P(F).$$

Proof

If the subsets representing these two events are also denoted by E and F, we have $E \subset F$. From the Venn diagram below (Fig. 11.3) we see that

number of elements in $E \le$ number of elements in F.

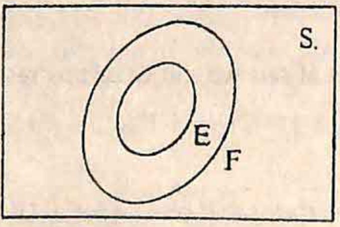


Fig. 11.3

Dividing both sides by the number of elements in the sample space S, we get

$$P(E) \le P(F)$$

which proves the theorem.

Example 11.6

A die is thrown twice. What is the probability that at least one of the two numbers is 4 ?

Solution

There are 36 possible outcomes which we may assume to be equally likely. The number of outcomes favourable to the occurrence of the event is 11. Thus, probability of the event is $\frac{11}{36}$

Example 11.7

In Example 11.6, let E denote the event: "the first die shows 4" and F the event: "the second die shows 4" Then $E \cap F$ is the event: "both dice show 4". Hence,

$$\begin{aligned} P\,(\text{at least one 4}) &= P\,(E \cup F\,) \\ &= P\,(E\,) + P\,(F\,) - P(E \cap F\,) \\ &= \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36} \end{aligned}$$

The same result as derived directly in Example 11.6.

EXERCISE 11.4

1. What is the probability that a number selected from the numbers 1, 2, ... 25 is a prime number ? You may assume that each of the 25 numbers is equally likely to be selected.
2. One card is drawn from a pack of 52 cards, each of the 52 cards being equally likely to be drawn. Find the probability of
   (a) the card drawn is red
   (b) the card drawn is a king
   (c) the card drawn is red and a king
   (d) the card drawn is either red or a king.
3. A bag contains 5 red balls, 3 black balls and 4 white balls. A ball is drawn out of the bag at random. What is the probability that the ball drawn is
   (i) white ?
   (ii) red ?
   (iii) black ?
   (iv) red or black ?

(v) red or white ?
(vi) red or white or black ?
(vii) yellow ?

4. A bag contains 100 identical tokens on which numbers 1 to 100 are marked. A token is drawn randomly. What is the probability that the number on the token is

(i) an even number ?
(ii) an odd number ?
(iii) a multiple of 3 ?
(iv) a multiple of 5 ?
(v) a multiple of 3 and 5 ?
(vi) a multiple of 3 or 5 ?
(vii) less then 20 ?
(viii) greater than 70 ?

5. A coin and a die are thrown. What is the probability of getting

(i) a head ?
(ii) an even number ?
(iii) a head and an even number ?
(iv) a head or an even number ?
(v) a tail and an odd number ?
(vi) a tail or an odd number ?

6. Say True or False giving reasons.

(i) $P(A) = \frac{1}{3}, P(B) = \frac{2}{3}$. A and B are mutually exclusive and exhaustive.

(ii) $P(A) = 0.4, P(B) = 0.25, P(A \text{ or } B) = 0.65$. A and B are mutually exclusive events.

(iii) $P(A) = 0.3, P(B) = 0.45, P(A \text{ and } B) = 0.2$. A and B are *not* mutually exclusive events.

(iv) $P(A) = 0.35, P(B) = 0.65$. A and B are complementary events.

11.8 Use of Permutations and Combinations in Calculation of Probabilities

To calculate the probability of an event, we have to count the total number of elementary events in the sample space and the number of elementary events favourable to the event. In simple cases the counting is easily done. However, in many cases direct counting is not easy. In such cases the use of permutations and combinations, which too are essentially based on counting operations, makes the calculation of probabilities easier. We illustrate the usefulness of this method by some examples.

Example 11.8

The random experiment consists of drawing four cards from a pack of 52 playing cards. Each group of 4 cards is an elementary event of this experiment, and the number of elementary events, which can all be considered to be equally likely, is equal to the numbers of sets of 4 cards that can be formed out of the 52 cards in the pack. Thus, the number of elementary events of the sample space of this experiment is given by $C(52, 4)$.

Consider now the event associated with this experiment which is "the 4 cards have the same value", that is, all four are aces, or kings, or queens, and so on. Then there are 13 elementary events which are favourable to the occurrence of this event. Hence, the probability of this event is

$$\frac{13}{C(52, 4)} = \frac{13 \times 1 \times 2 \times 3 \times 4}{52 \times 51 \times 50 \times 49} = 0.000048 \text{ (approx.)}.$$

Consider next the event that "all 4 cards are of the same colour", that is, either all are red or all are black. There are 26 red cards from which 4 can be selected in $C(26, 4)$ ways. Similarly, out of 26 black cards also 4 cards can be selected in $C(26, 4)$ ways. Thus, out of all groups of cards we will have $C(26, 4)$ groups having only red cards and $C(26, 4)$ groups having only black cards. Therefore, the number of elementary events favourable to the occurrence of this event is given by

$$C(26, 4) + C(26, 4) = \frac{2 \times 26 \times 25 \times 24 \times 23}{1 \times 2 \times 3 \times 4}$$

Hence, the probability of this event is given by

$$\frac{2 \times 26 \times 25 \times 24 \times 23}{52 \times 51 \times 50 \times 49} = 0.11 \text{ (approx.)}.$$

Example 11.9

The random experiment consists of first selecting three numbers out of the numbers 1, 2, 3, 4 and 5 and then writing down all possible arrangements of these numbers. For example, if numbers 1, 2 and 5 are selected, we will get 6 arrangements (1, 2, 5), (2, 1, 5), (1, 5, 2), (5, 1, 2), (2, 5, 1) and (5, 2, 1). Three numbers out of the given five can be selected in $C(5, 3)$ ways, and each set of 3 numbers gives rise to $3! = 6$ arrangements. Thus, the total number of outcomes is

$$C(5, 3) \times 3! = P(5, 3) = 60.$$

Each of the $C(5, 3)$ sets of 3 numbers gives rise to only one arrangement in which the numbers are in the natural order. For example, if the 3 selected numbers are 2, 1 and 4 they will result in $3! = 6$ arrangements out of which only one arrangement 1, 2, 4 will have the three numbers in natural order. Hence, for the event "the arranged numbers are in natural order" the number of favourable cases is $C(5, 3)$ so that, the probability of this event is

$$\frac{C(5, 3)}{P(5, 3)} = \frac{1}{3!} = \frac{1}{6} = 0.17 \text{ (approx.)}.$$

You will have noted that in these examples the knowledge of permutations and combinations has made the calculation of the number of favourable cases and the number of equally likely cases quite easy.

EXERCISE 11.5

1 A class consists of 10 boys and 8 girls. A committee of 3 students is constituted. What is the probability that the committee has

(i) all boys ?
(ii) all girls ?
(iii) 1 boy and 2 girls ?
(iv) at least one girl ?
(v) atmost one girl ?

2. An urn contains 4 black balls, 3 white balls and 5 red balls. Two balls are drawn at random. What is the probability that the balls are

(i) red ?
(ii) black ?
(iii) white ?
(iv) 1 black and 1 white ?
(v) 1 white and 1 red ?
(vi) not red ?
(vii) not white ?

3. Two cards are drawn together from a pack of 52 cards at random. What is the probability that

(i) both are spades ?
(ii) one is a spade and one is a heart ?
(iii) both are kings ?
(iv) exactly one is a king ?

4. Three light bulbs are selected at random from 20 bulbs of which 5 are defective. What is the probability that

(i) none of the bulbs is defective ?
(ii) exactly one is defective ?
(iii) at least one is defective ?

5. Two cards are drawn at random from 8 cards numbered from 1 to 8. What is the probability that the sum of the numbers is odd, if the two cards are drawn together ?

11.9 General Definition of a Random Experiment

The concept of a random experiment was introduced in Sec. 11.2, and a definition of probabilities of events associated with a random experiment was given in Sec. 11.6 according to which each elementary event of the sample space had probability $\frac{1}{n}$, n being the total number of elementary events in the sample space.

Instead of defining the probability of a general event E by

$$P(E) = \frac{\text{no. of elementary events favourable to } E}{\text{total no. of elementary events in the sample space}}$$

we could as well have begun by defining the probability of each elementary event as $\frac{1}{n}$. Then, if E is the general event represented by the set $\{\omega_1, \ldots, \omega_k\}$ of elementary events, we have by the addition rule of probability $P(E) = \frac{k}{n}$, since the elementary events are mutually exclusive (See corollary to Theorem 11.1).

Thus, the assignment of probabilities to events by means of the definition of Sec. 11.6 is really based on the assumption that the elementary events are equally likely to occur and have, therefore, equal probabilities.

In the case of some random experiments, the assumption that the elementary events are equally likely to occur can be accepted. When a coin is tossed we have no reason to doubt that heads and tails are equally likely outcomes so that the probability of occurrence of either of them could be taken to be $\frac{1}{2}$. This assumption is an expression of our belief that the coin is not a "crooked" or a "biased", coin. Our knowledge of physics would tell us that if the coin is not of uniform density but has a higher density towards one of the faces, then the two faces cannot be considered to be equally likely to appear at the throw of the coin. In fact, the coin is almost sure to show one face only (which one ?) almost all the time. In such a case it would not be proper to regard the two elementary events as equally likely.

There are other situations where too the assumption that the elementary events are equally likely leads to difficulties and is not acceptable. For example, imagine an experiment which is carried out in two stages.

First stage : toss a coin
Second stage : if the coin shows head toss it again but if it shows tail then toss a die.

We will assume for our discussion that the elementary events of the experiment of tossing a coin are equally likely, and that the same holds for the six outcomes of tossing a die.

If the first stage results in a head, the second stage will have a head or a tail as the outcome. If, on the other hand, the first stage outcome is a tail, the second stage-outcomes will be any of the numbers 1, 2, 3, 4, 5 and 6. Thus, there are eight possible outcomes of this experiment, and the sample space can be described by the set

$$\{(H, H), (H, T), (T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\}.$$

Consider the event represented by the set $\{(H, H), (H, T)\}$. If the first stage results in head, then, and only then, we will observe the occurrence of this event. So, this set can be taken to represent the event: "first stage results in a head". Similarly, other events associated with the first or the second stage only, can also be represented by subsets of the sample space. For example, the set $\{(H, T)\}$ represents the event: "second stage results in a tail", the set $\{(T, 2), (T, 4), (T, 6)\}$ represents the event: "second stage results in an even number", and so on.

In this random experiment, we cannot consider the eight elementary events as equally likely. If we do, then the probability of the event represented by the subset $\{(H, H), (H, T)\}$ would be equal to $\frac{2}{8} = \frac{1}{4}$ But this subset can also be taken to represent the event that the throw of the coin at the first stage results in head and, therefore, its probability should be equal to $\frac{1}{2}$. Similarly, the subset

$$\{(T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\}$$

represents an event with probability $\frac{6}{8} = \frac{3}{4}$ if we regard the eight outcomes of the experiment as equally likely. At the same time, it can be taken to represent the event that the throw of the coin at the first stage results in tail and should, therefore, have probability equal to $\frac{1}{2}$.

We have seen that the eight outcomes of our experiment cannot be considered equally likely since that assumption leads to difficulties. How do we attach probabilities to the events associated with this experiment ? We will begin with the elementary events and see what should be the probabilities attached to them.

From our disussion above it seems reasonable to assume that

$$P(\{(H, H), (H, T)\}) = \frac{1}{2}$$

and

$$P(\{(T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\}) = \frac{1}{2}$$

If the probabilities satisfy the addition rule we have

$$P(\{(H, H)\}) + P(\{(H, T)\}) = \frac{1}{2}$$

If the first throw of the coin results in head, the coin is thrown again and the two outcomes H and T are, according to the assumption made in the beginning, equally likely. Therefore, we must have

$$P(\{(H, H)\}) = P(\{(H, T)\}) = \frac{1}{4}.$$

Similarly, as we have stated earlier, it is reasonable to assume

$$P(\{(T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\}) = \frac{1}{2},$$

since it is the probability of the event that the first stage results in a tail. If the first throw of the coin results in a tail, the six outcomes of throwing a die are equally likely. So, we should have

$$\begin{aligned} P(\{(T, 1)\}) &= P(\{(T, 2)\}) = P(\{(T, 3)\}) = P(\{(T, 4)\}) \\ &= P(\{(T, 5)\}) = P(\{(T, 6)\}) \end{aligned}$$

which means that each of these probabilities is equal to $\frac{1}{12}$.

So, in the random experiment described above we see that the probabilities of the elementary events, which could not be considered to be equally likely, cannot be taken as equal. A more reasonable way is to take the probabilities of the elementary events

$$\{(H, H)\}, \{(H, T)\}, \{(T, 1)\}, \{(T, 2)\}, \{(T, 3)\}, \{(T, 4)\}, \{(T, 5)\}, \{(T, 6)\}$$

equal to $\frac{1}{4}, \frac{1}{4}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}$ respectively.

Thus, we see that merely describing the sample space (the set of outcomes) is really not enough to describe a random experiment. If the elementary events can be considered equally likely, the sample space S describes the random experiment completely, in the sense that knowing S one can calculate the probabilities of all events associated with the random experiment. But if the elementary events cannot be considered to be equally likely, then the set S does not completely describe the random experiment, because the probabilities of the events associated with the experiment cannot be calculated by the knowledge of S alone.

How does one then define the probability of events in those cases where the elementary events of the random experiment cannot be considered equally likely ? The way out is to select positive numbers (that is, number > 0) $p_1, p_2, \ldots, p_n$ such that $\sum_{i=1}^{n} p_i = 1$, and to assign these numbers in a one-to-one manner to the n elementary events of the sample space of the random experiment. We call these numbers the probabilities of occurrence of the elementary events. The values selected for $p_1, p_2, \ldots, p_n$ reflect our understanding of the experiment, a higher value being given to the probability assigned to an elementary event which we consider more likely to occur. In the example discussed above the positive numbers were $\frac{1}{4}, \frac{1}{4}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}$, whose sum, we see, is equal to unity. The actual values of the

S.C.E.R.T. W.B. LIBRARY
9934
16.5.2k

probability assigned to any particular elementary event was obtained by analysing the nature of the experiment. If we have reason to believe that the n elementary events forming the sample space are equally likely, we will take $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$, so that the equally likely case is also covered by the general procedure.

Consider now an event E associated with the sample space of the experiment in which the n elementary events forming the sample space have been assigned probabilities $p_1, p_2, \ldots, p_n$. The event E is represented by a subset of the sample space and we define the probability of occurrence of E as the sum of the values p_i assigned to the elementary events belonging to the subset representing the event. In our example, the event "the throw of the die at the second stage results in an even number" is presented by the subset $\{(T, 2), (T, 4), (T, 6)\}$ and would, thus, have probability equal to $\frac{1}{12}+\frac{1}{12}+\frac{1}{12}=\frac{1}{4}$.

Similarly, the event "head at first throw, tail at second throw" is represented by the single point set $\{(H, T)\}$ and will have probability equal to $\frac{1}{4}$.

In general, suppose $S = \{\omega_1, \omega_2, \ldots, \omega_n\}$ is the sample space of the random experiment and probabilities $p_1, p_2, \ldots, p_n$ are assigned to the elementary events $\omega_1, \omega_2, \ldots, \omega_n$ respectively. Then, the probability $P(E)$ of any event represented by the subset E of S can be defined as

$$P(E) = \sum_{\omega_i \in E} p_i.$$

With this definition, we have $P(\phi) = 0$, $P(S) = 1$ and for all other events E the probability lies between 0 and 1. Not only that, all the theorems proved in Sec. 11.7 still hold if $P(E)$ is defined in the manner given here starting with the values $p_1, p_2, \ldots, p_n$. With this general definition we will have

(i) $P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$

$= P(E) + P(F)$ for mutually exclusive events,

(ii) $P(\text{not-}E) = 1 - P(E)$,

(iii) $P(E) \le P(F)$ if E implies F

If we take each $p_i = \frac{1}{n}$, then we get our earlier definition, namely

$$P(E) = \frac{\text{No. of favourable elementary events}}{\text{Total no. of elementary events}}$$

We are now in a position to give the mathematical definition of random experiment. The definition is as follows:

Definition

A random experiment is defined by the set $\mathbf{S} = \{\omega_1, \omega_2, \ldots, \omega_n\}$ of its outcomes, to each outcome ω_i of which is associated a positive number p_i, called the probability of ω_i, such that $p_1 + p_2 + \ldots + p_n = 1$. The probability $P(E)$ of any event associated with the random experiment is the sum of the probabilities of the elementary events contained in the subset of $\mathbf{S}$ representing the event.

EXERCISE 11.6

1. A ball is drawn from an urn containing one red ball and one black ball. If the ball drawn is red a coin is tossed, if it is black a die is thrown. What is the probability of
   - (i) each elementary event ?
   - (ii) getting a head ?
   - (iii) getting an even number ?

2. A ball is drawn from an urn containing 2 red balls and 3 white balls. If the ball is red, a card from a pack of playing cards is selected and if the ball is white, two coins are tossed together. What is the probability of
   - (i) each elementary event ?
   - (ii) getting a spade ?
   - (iii) getting a spade or a heart ?
   - (iv) getting two heads ?
   - (v) getting exactly one tail ?

B. CONDITIONAL PROBABILITY AND INDEPENDENCE

11.10 Conditional Probability

We shall first discuss the concept of conditional probability in the context of those random experiments only in which the outcomes can be considered as equally likely to occur when the experiment is performed. Consider, for example, the random experiment of drawing a ball from a box containing 20 red, 30 black, and 50 white balls. The balls differ only in their colour and are otherwise identical in all respects. We can, therefore, regard the elementary events of the experiment as being equally likely.

The probability of the event "the ball drawn is red" is equal to $\frac{20}{100} = 0.2$. Now, suppose we are told, after the ball has been drawn, that it is not white (that is, it is either red or black). Should we, in view of this additional information, revise the probability of the ball drawn being red, or leave the value of the probability unchanged ? If we had been told that the ball drawn was neither black nor white we would be certain that it was red. In this case, we would agree that the probability of the event "the ball drawn is red" should, in view of the additional information available, be taken as 1. Or, suppose the box has 20 red, 1 black and 79 white balls. The probability of drawing a red ball from this box is the same as before, i.e. 0.2. However, if we are told in this case that the ball drawn is not white, we would give a very high probability to the ball being red, since we now know that the ball drawn is one of the 21 red or black balls out of which 20 are red. We thus see that if we are told, after the experiment has been performed, that a particular event has occurred, then we are led to revise the value of the probabilities of the other events in the light of the additional information available

Let us go back to the example of the box with 20 red, 30 black, and 50 white balls. A ball is drawn and we are told that it is not white. With this information we are sure that the ball is either red or black. So, the probability of the ball being red may now be taken as equal to the probability of drawing a red ball from a box containing 20 red and 30 black balls. In other words, the additional information really amounts to telling us that the situation may be considered as being that of a new random experiment for which the sample space consists of all those outcomes only which are favourable to the occurrence of the event "the ball drawn is not white". In this new random experiment the probability of drawing a red ball is given by

$$\frac{\text{No. of red balls}}{\text{No. of red balls} + \text{No. of black balls}} = \frac{20}{50} = 0.4.$$

Thus, the information that the ball drawn is not white has changed the probability of the ball being red from 0.2 to 0.4. We call this changed probability the *conditional probability* of drawing a red ball given that the ball drawn is not white. It is called conditional because it has been obtained under the condition that the experiment had resulted in the event "the ball drawn is not white".

Thus, in general we can talk of the conditional probability of the occurrence of an event E, given that an event F is known to have occurred. We shall denote this conditional probability by $P(E/F)$ where E and F are events associated with the same sample space. To calculate the value of $P(E/F)$, we take the elementary events favourable to the occurrence of F as our new sample space, and then find out how many elementary events out of these are favourable to the occurrence of E. We have then

$$P(E/F) = \frac{\begin{array}{c}\text{No. of elementary events favourable to } F\\ \text{which are also favourable to } E\end{array}}{\text{No. of elementary events favourable to } F}$$

For example, suppose the experiment is of drawing a card from a pack of playing cards, and the events F and E are respectively "the card drawn is red" and "the card drawn is an ace". Then there are 26 elementary events favourable to F. Out of the 26 elementary events favourable to F, there are 2 which are favourable to E. The subset consisting of these two elementary events represents the event "E and F". Thus, we may write the conditional probability $P(E/F)$ as

$$P(E/F) = \frac{\text{No. of elementary events favourable to the occurrence of the event "}E\text{ and }F\text{"}}{\text{No. of elementary events favourable to the occurrence of the event }F}$$

Dividing the numerator and denominator by the total number of elementary events of the experiment, we see that $P(E/F)$ can also be written as

$$P(E/F) = \frac{P(E \text{ and } F)}{P(F)}$$

We use this form for the *definition* of conditional probability.

Definition

Given two events E and F associated with the sample space of the same random experiment, the conditional probability $P(E/F)$ of the occurence of E knowing that event F has occurred is given by

$$P(E/F) = \frac{P(E \text{ and } F)}{P(F)} = \frac{P(E \cap F)}{P(F)}$$

Note that we are assuming that $P(F) \neq 0$; if $P(F) = 0$, the expression $\frac{P(E \cap F)}{P(F)}$ makes no sense.

We arrived at this definition after a discussion in which the elementary events of the experiment were considered to be equally likely and the corresponding definition of the probability of an event was used. However, the same definition of conditional probability can also be used in the general case where the elementary events are not equally likely, the probabilities $P(E \text{ and } F)$ and $P(F)$ being calculated according to the general method.

Example 11.10

Consider the experiment described in Sec. 11.9 with the sample space

$$\{(H, H), (H, T), (T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\}.$$

The probabilities assigned to these 8 elementary events were $\frac{1}{4}, \frac{1}{4}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}$, respectively. Let F be the event that the first throw of the coin results in a tail, and E the event that the die shows a number greater than 4. Then

$$P(E) = P(\{(T, 5), (T, 6)\}) = \frac{1}{12} + \frac{1}{12} = \frac{1}{6}$$

$$P(F) = P(\{(T, 1), (T, 2\}, (\mathrm{T}, 3), (\mathrm{T}, 4), (\mathrm{T}, 5), (T, 6)\}) = \frac{1}{2}$$

$$P(E \text{ and } F) = P(\{(T, 5), (T, 6)\}) = \frac{1}{6}\cdot$$

Hence,

$$P(E/F) = \frac{P(E \cap F)}{P(F)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

Example 11.11

A die is thrown twice and the sum of the numbers appearing is observed to be 6. What is the conditional probability that the number 4 has appeared at least once ?

Solution

If E is the event : "number 4 appears at least once" and F the event : "the sum of the numbers appearing is 6", then we have, assuming the 36 outcomes to be equally likely,

$$P(E) = \frac{11}{36} \text{ and } P(F) = \frac{5}{36},$$

since the elementary events favourable to the occurrence of E are (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (1, 4), (2, 4), (3, 4), (5, 4), (6, 4) and those favourable to the occurrence of F are (1, 5), (2, 4), (3, 3), (4, 2), (5, 1). Also, the elementary events favourable to the occurrence of both E and F are (2, 4) and (4, 2) so that we have

$$P(E \text{ and } F) = P(E \cap F) = \frac{2}{36}\cdot$$

Hence, the probability we are required to find is

$$P(E/F) = \frac{P(E \cap F)}{P(F)} = \frac{2}{5}\cdot$$

EXERCISE 11.7

1. A coin is tossed twice and the four possible outcomes are assumed to be equally likely. If E is the event: "both head and tail have occurred", and F the event: "at most one tail has occurred", find $P(E)$, $P(F)$, $P(E/F)$ and $P(F/E)$.

2. A bag contains 3 red and 4 black balls and another bag has 4 red and 2 black balls. One bag is selected, each of the two bags being equally likely to be selected. From the

selected bag a ball is drawn, each ball in the bag being equally likely to be drawn. Let E be the event: "the first bag is selected", F the event: "the second bag is selected" and G the event: "the ball drawn is red". Find $P(E)$, $P(F)$, $P(G/E)$, $P(G/F)$.

11.11 Independent Events

Consider the experiment of drawing a card from a pack of 52 playing cards, in which the elementary events can be supposed to be equally likely. If E and F denote the events "the card drawn is a spade" and "the card drawn is an ace" respectively, then

$$P(E)=\frac{\text{No. of spade cards}}{52}=\frac{13}{52}=\frac{1}{4}$$

and

$$P(F)=\frac{\text{No. of aces}}{52}=\frac{4}{52}=\frac{1}{13}\cdot$$

Also, "E and F" is the event "the card drawn is the ace of spades", so that

$$P(E \text{ and } F)=\frac{1}{52}\cdot$$

Hence, in this case

$$P(E/F)=\frac{P(E\cap F)}{P(F)}$$

$$=\frac{\frac{1}{52}}{\frac{1}{13}}=\frac{1}{4}=P(E),$$

so that $P(E)$ and $P(E/F)$ are equal. This shows that the information that the event F has occurred has not changed the probability of occurrence of the event E. Not only that, we also have

$$P(F/E)=\frac{P(E\cap F)}{P(E)}$$

$$=\frac{\frac{1}{52}}{\frac{1}{4}}=\frac{1}{13}=P(F),$$

so that the information that E has occurred does not change the probability of occurrence of the event F.

Two events which are such that the information that one of them has occurred does not change the probability of occurrence of the other are called *independent events*. The exact definition is as follows:

Definition

Two events E and F defined on the sample space **S** of a random experiment are said to be *independent* if

$$P(E \text{ and } F) = P(E) \cdot P(F).$$

This definition can be used even for events associated with a random experiment in which the elementary events are not equally likely. The effect of independence of the events E and F on the conditional probabilities $P(E/F)$ and $P(F/E)$ is given by the theorem below.

Theorem 11.5

If the events E and F defined on the sample space **S** of a random experiment are independent, then

$$P(E/F) = P(E)$$

and

$$P(F/E) = P(F).$$

Proof

Since E and F are independent events, we have

$$P(E \text{ and } F) = P(E) \cdot P(F)$$

Hence,

$$P(E/F) = \frac{P(E \cap F)}{P(F)}$$

$$= \frac{P(E) \cdot P(F)}{P(F)} = P(E).$$

Similarly,

$$P(F/E) = \frac{P(E \cap F)}{P(E)}$$

$$= \frac{P(E) \cdot P(F)}{P(E)} = P(F)$$

which proves the theorem.

When E and F are independent events the probability $P(E \text{ and } F)$ is the product of the probabilities $P(E)$ and $P(F)$. In the general case, where E and F may or may not be independent, we have

$$\begin{aligned} P(E \text{ and } F) &= P(E) \cdot P(F/E) \\ &= P(F) \cdot P(E/F) \end{aligned}$$

This expression, sometimes called the *multiplication rule* of probabilities, is simply a restatement of the definition of $P(E/F)$ and $P(F/E)$.

It should be noted that the notion of independent events is nothing more than a statement about the probability $P(E \text{ and } F)$ being the product of the probabilities $P(E)$ and $P(F)$. We should not read anything more than this in the words "independent events". A physical description of the events cannot tell us if they are independent or not. Only when the probabilities $P(E \text{ and } F)$, $P(E)$ and $P(F)$ have been calculated can we say anything about their independence?

Example 11.12

A die is thrown and the 6 possible outcomes are assumed to be equally likely. If E is the event: "the number appearing is a multiple of 3", and F the event: "the number appearing is even", we have

$$P(E) = P(\{3, 6\}) = \frac{1}{3}$$

$$P(F) = P(\{2, 4, 6\}) = \frac{1}{2}$$

$$P(E \cap F) = P(\{6\}) = \frac{1}{6} = P(E) \cdot P(F).$$

Hence, events E and F are independent.

Example 11.13

In the sample space of Section 11.9, let E be the event: "the second throw of the coin results in head or the throw of die results in 3", and F the event: "the throw of die results in an odd number". Then

$$P(E) = P(\{(H, H), (T, 3)\}) = \frac{1}{4} + \frac{1}{12} = \frac{1}{3}$$

$$P(F) = P(\{(T, 1), (T, 3), (T, 5)\}) = \frac{1}{12} + \frac{1}{12} + \frac{1}{12} = \frac{1}{4}$$

$$P(E \cap F) = P(\{(T, 3)\}) = \frac{1}{12} = P(E) \cdot P(F).$$

Thus, events E and F are independent.

EXERCISE 11.8

1. One card is drawn from a pack of 52 cards so that each card is equally likely to be selected. In which of the following cases are the events E and F independent ?
   (a) E: "the card drawn is a spade"
   F: "the card drawn is an ace"
   (b) E: "the card drawn is black"
   F: "the card drawn is a king"
   (c) E: "the card drawn is a king or queen"
   F: "the card drawn is a queen or jack"
2. A coin is tossed thrice and all eight outcomes are assumed equally likely. In which of the following cases are the events E and F independent ?
   (a) E: "the first throw results in head"
   F: "the last throw results in tail"
   (b) E: "the number of heads is two"
   F: "the last throw results in head"
   (c) E: "the number of heads is odd"
   F: "the number of tails is odd"

11.12 Independent Experiments

Consider two random experiments: one consists in tossing a coin and has two equally likely outcomes, the other in throwing a die with six equally likely outcomes. The sample space of the first experiment may be described by the set $\{H, T\}$, and that of the second by the set $\{1, 2, 3, 4, 5, 6\}$. We may regard the two experiments together as forming a new random experiment with a sample space of 12 equally likely outcomes described by the set

$$\{(H, 1), (H, 2), \ldots, (H, 6), (T, 1), (T, 2), \ldots, (T, 6)\}.$$

The event E: "the coin shows head" is represented by the subset $\{H\}$ in the sample space of the first experiment and has probability $\frac{1}{2}$. The event E can also be considered as an event associated with the combined experiment and is then represented by the subset

$$\{(H, 1), (H, 2), (H, 3), (H, 4), (H, 5), (H, 6)\}$$

and has probability equal to $\frac{1}{2}$ in the combined experiment also.

Similarly, the event F: "the number on the die is divisible by 3" is represented by the subset $\{3, 6\}$ in the sample space of the second experiment, and by the subset $\{(H, 3), (H, 6), (T, 3), (T, 6)\}$ in the sample space of the combined experiment. And in both cases it has probability $\frac{1}{3}$

From these two events E and F we can also define the event "E and F" in the sample space of the combined experiment. It is represented by the subset $\{(H, 3), (H, 6)\}$ and therefore, has probability $\frac{2}{12} = \frac{1}{6}$. Thus, we find that for the events E and F as defined above we have

$$P(E \text{ and } F) = P(E) \cdot P(F).$$

Thus, the events E and F may be called independent. A similar result will hold for any two events if one of them is defined in terms of the first experiment, and the other in terms of the second experiment only. For example, if E is the event: "the coin shows head", and F the event: "the number on the die is less than 4", then we have

$$P(E) = \frac{1}{2}, \quad P(F) = \frac{3}{6} = \frac{1}{2}$$

and

$$P(E \text{ and } F) = P(\{(H, 1), (H, 2), (H, 3)\})$$
$$= \frac{3}{12} = \frac{1}{4} = P(E) \cdot P(F).$$

But if the events E and F are not defined in terms of the first and second events alone, such a result may not hold. For example, let E be the event: "the coin shows head and the die shows a number more than 4", and F the event: "the coin shows head and the die shows an even number" Then

$$P(E) = P(\{(H, 5), (H, 6)\}) = \frac{2}{12} = \frac{1}{6}$$

$$P(F) = P(\{(H, 2), (H, 4), (H, 6)\}) = \frac{3}{12} = \frac{1}{4}$$

and

$$P(E \text{ and } F) = P(\{(H, 6)\}) = \frac{1}{12} \neq P(E) \cdot P(F).$$

Thus, in our example, if the event "E and F" is interpreted as "E occurs in the first and F in the second experiment", then we have

$$P(E \text{ and } F) = P(E) \cdot P(F)$$

so that we can say that events E and F are independent. Since this happens for all pairs of events E and F, where E is associated with the first and F with the second experiment, we say that the two experiments are independent. The exact definition of independent experiments is as follows:

Definition

Two random experiments are said to be independent if, for every pair of events E and F where E is associated with the first and F with the second experiment, the probability of the simultaneous occurrence of the events E and F, when the two experiments are performed, is the product of the probabilities $P(E)$ and $P(F)$ calculated separately on the basis of the two experiments.

Consider next a combined experiment which is as follows. A and B denote two families with A having 3 children of which 1 is a boy and 2 are girls, and B having 2 children of which 1 is a boy and other a girl. We first select a family and assume that each is equally likely to be selected. Then from the selected family we select a child, again assuming that every child of the family is equally likely to be selected.

The sample space of the combined experiment may be taken as the set

$$\{Ab_1, Ag_1, Ag_2, Bb_2, Bg_3\}.$$

Here we have used g_1, g_2 to indicate the two girls belonging to family A, and g_3 to indicate the girl belonging to family B. Similarly, b_1 and b_2 have been used to denote boys belonging to families A and B respectively. The elementary events in this case are, however, not equally likely. The subset $\{Ab_1, Ag_1, Ag_2\}$ represents the events that family A is selected at first, and should, therefore, have probability equal to $\frac{1}{2}$. Similarly, the subset $\{Bb_2, Bg_3\}$ should also have probability $\frac{1}{2}$. Once family A is selected each of the three children has the same chance of being selected. So we should have

$$P(\{Ab_1\}) = P(\{Ag_1\}) = P(\{Ag_2\}) = \frac{1}{6}.$$

Similarly,

$$P(\{Bb_2\}) = P(\{Bg_3\}) = \frac{1}{4}.$$

Hence, the combined experiment is defined by the sample space

$$\{Ab_1, Ag_1, Ag_2, Bb_2, Bg_3\}$$

with the elementary events having respectively the probabilities $\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{4}, \frac{1}{4}$. Let now E be the event: "family A is selected at first", and F the event: "the selected child is a girl". We may say that E is defined in terms of the first experiment and F in terms of the second. We have

$$P(E) = P(\{Ab_1, Ag_1, Ag_2\}) = \frac{1}{2},$$

$$P(F) = P(\{Ag_1, Ag_2, Bg_3\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{4} = \frac{7}{12},$$

$$P(E \text{ and } F) = P(\{Ag_1, Ag_2\}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

so that

$$P(E \text{ and } F) \neq P(E) \cdot P(F).$$

Hence, in this case we cannot say that the two experiments of selecting a family and then selecting a child are independent experiments.

If we examine the two examples closely we notice an important difference. In the first example the sample space of the second experiment is the same for all outcomes of the first experiment: it is {1, 2, 3, 4, 5, 6} irrespective of whether the first experiment results in head or tail. The nature of the two experiments is such that we can state without any hesitation that the occurrence of an event in the first experiment does not in any way affect the occurrence of any event in the second experiment. In other words we could say that the probability of occurrence of any event in the second experiment is the same no matter what the outcome of the first experiment. It is due to this that we get the result

$$P(E \text{ in first and } F \text{ in second})$$

$$= P(E \text{ in first}) \cdot P(F \text{ in second})$$

for all events E, F with E belonging to the first and F to the second experiment.

But the situation is different in the second example in which, if the first experiment results in the selection of family A, the sample space of the second experiment is {boy, girl, girl}; but, if family B is selected in the first experiment the sample space of the second experiment is {boy, girl}. Since the sample space of the second experiment depends on the outcome of the first experiment, we cannot assert that the outcome of the first experiment will have no effect on the probabilities of occurrence of different events in the second experiment. In fact, the probability of selecting a boy in the second experiment is $\frac{1}{3}$ if the first experiment results in the selection of family A, but it becomes equal to $\frac{1}{2}$ if family B is selected in the first experiment.

It should be noted that when we are dealing with two events E and F associated with the same experiment the description of the events is not sufficient to enable us to decide if E and F are independent events or not. The decision about the independence of the two events can only be made after the probabilities P (E), $P(F)$, and $P(E$ and $F)$ have been calculated. But the situation is different when we are dealing with two experiments. We can in many cases conclude that the experiments are independent or not from the physical description of the experiments without having to do any probability calculations. What is more, if we can say on the basis of the description of the experiments that they are independent, i.e. the occurrence of an event in one experiment has no effect on the probability of occurrence of an event in the second experiment, then we can use this knowledge to calculate the probabilities

$$P\,(E \text{ in first experiment and } F \text{ in second experiment})$$

as the products of the probabilities

$$P\,(E \text{ in first experiment}) \text{ and } P\,(F \text{ in second experiment})$$

In general, suppose two random experiments with sample spaces S_1 and S_2 are independent experiments. Let the elementary events of S_1 be $x_1, x_2, \ldots, x_m$ with probabilities $p_1, p_2, \ldots, p_m$ respectively and let the elementary events of S_2 be $y_1, y_2, \ldots, y_n$ with probabilities $q_1, q_2, \ldots, q_n$. Then the combined experiment has mn elementary events

$$(x_i, y_j\,),\ i = 1, \ldots, m;\ j = 1, \ldots, n$$

with probabilities $p_i \cdot q_j$.

Now suppose E is an event associated with the first experiment represented by the subset $\{x_{i1}, x_{i2}, \ldots, x_{ir}\}$ of S_1 and F an event associated with the second experiment represented by the subset $\{y_{j1}, y_{j2}, \ldots, y_{js}\}$ of S_2. Then the event "E and F" associated with the combined experiment is represented by the subset

$$\begin{aligned}
\{\ &(x_{i1}, y_{j1}), (x_{i1}, y_{j2}), \ldots, (x_{i1}, y_{js}\,),\\
&(x_{i2}, y_{j1}), (x_{i2}, y_{j2}), \ldots, (x_{i2}, y_{js}\,),\\
&\ldots, \ldots, \ldots\\
&(x_{ir}, y_{j1}), (x_{ir}, y_{j2}), \ldots\ (x_{ir}, y_{js}\,)\ \}
\end{aligned}$$

of rs elementary events of the combined experiment. According to the general definition of a random experiment, we have

$$P\,(E\,) = p_{i1} + p_{i2} + \ldots + p_{ir}$$
$$P\,(F\,) = q_{j1} + q_{j2} + \ldots + q_{js}$$

and

$$\begin{aligned}
P\,(E \text{ and } F\,) &= p_{i1}\,q_{j1} + p_{i1}\,q_{j2} + \ldots + p_{i1}\,q_{js} +\\
&\quad\ p_{i2}\,q_{j1} + p_{i2}\,q_{j2} + \ldots + p_{i2}\,q_{js} +\\
&\quad\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\quad\ p_{ir}\,q_{j1} + p_{ir}\,q_{j2} + \ldots + p_{ir}\,q_{js}\\
&= (p_{i1} + p_{i2} + \ldots + p_{ir}\,)\,(q_{j1} + q_{j2} + \ldots + q_{js}\,)\\
&= P\,(E\,) \cdot P\,(F\,)
\end{aligned}$$

which confirms that two random experiments are indeed independent.

Example 11.14

One ball is drawn from a bag containing 3 red and 2 black balls. Its colour is noted and then it is put back in the bag. A second draw is made and the same procedure is repeated. At each draw the probability of drawing a red ball is $\frac{3}{5}$ and of drawing a black ball is $\frac{2}{5}$, assuming that each ball has the same chance of being drawn. The two experiments are independent as the result of the first draw has no effect on the result of the second draw. Hence,

$$P\,(\text{2 red balls are drawn}) = \frac{3}{5}\cdot\frac{3}{5} = \frac{9}{25}$$

$$P\,(\text{the balls drawn are of different colour}) = \frac{3}{5}\cdot\frac{2}{5} + \frac{2}{5}\cdot\frac{3}{5} = \frac{12}{25}$$

$$P\,(\text{both balls drawn are black}) = \frac{2}{5}\cdot\frac{2}{5} = \frac{4}{25}\,.$$

Example 11.15

A bag has 4 red and 5 black balls, a second bag has 3 red and 7 black balls. One ball is drawn from the first and two from the second. The possible outcomes at each draw are assumed to be equally likely. At the first draw, we have

$$P\,(\text{red}) = \frac{4}{9},\quad P\,(\text{black}) = \frac{5}{9}.$$

At the second draw there are $C\,(10, 2)$ possible outcomes of which $C\,(3, 2)$ have both balls red, $C\,(3, 1)\cdot C\,(7, 1)$ have one ball red and one black, and $C\,(7, 2)$ have both balls black. Hence, for the second draw, we have

$$P\,(\text{red, red}) = \frac{C\,(3, 2)}{C\,(10, 2)} = \frac{1}{15}$$

$$P\,(\text{red, black}) = \frac{C\,(3, 1)\cdot C\,(7, 1)}{C\,(10, 2)} = \frac{7}{15}$$

$$P\,(\text{black, black}) = \frac{C\,(7, 2)}{C\,(10, 2)} = \frac{7}{15}\,.$$

We can consider the two draws as forming two independent events. Hence, we get

$$P\text{ (red at first draw, red and black at second draw)} = \frac{4}{9}\cdot\frac{7}{15} = \frac{28}{135}$$

$$P\text{ (black at first draw, two reds at second draw)} = \frac{5}{9}\cdot\frac{1}{15} = \frac{1}{27}.$$

Consider now the event: "two blacks and a red" in the combined experiment. This event occurs when the following mutually exclusive events occur: "red at first draw, two blacks at second draw", "black at first draw, red and black at second draw". Hence,

$$P\text{ (two blacks and a red)} = \frac{4}{9}\cdot\frac{7}{15} + \frac{5}{9}\cdot\frac{7}{15}$$

$$= \frac{7}{15}.$$

Similarly,

$$P\text{ (two reds and a black)} = \frac{4}{9}\cdot\frac{7}{15} + \frac{5}{9}\cdot\frac{1}{15}$$

$$= \frac{11}{45}.$$

EXERCISE 11.9

1. In Example 11.15 find P (all three balls are red), P (all three balls are black).
2. A coin is tossed three times and all possible outcomes are considered equally likely. Regarding this as a combined experiment with tossing a coin twice as first and tossing a coin once as the second experiment, the two experiments being independent, calculate the following probabilities:
   (a) P (two heads and a tail)
   (b) P (three heads).

   Compare these probabilities with those obtained directly by considering the throwing of a coin three times as a single experiment.

11.13 The Law of Total Probability and Bayes' Formula

Let S be the sample space and let H_1, H_2 be two mutually exclusive and exhaustive events (Fig. 11.4). Let E be another event that occur with H_1 or H_2.

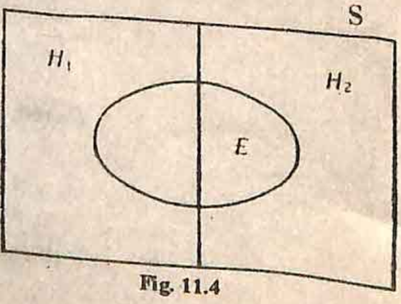


Fig. 11.4

Now $S = H_1 \cup H_2$

$$S \cap E = (H_1 \cup H_2) \cap E = (H_1 \cap E) \cup (H_2 \cap E)$$

$H_1 \cap E$ and $H_2 \cap E$ are mutually exclusive.

Hence,
$$\begin{aligned} P(E) = P(S \cap E) &= P[(H_1 \cap E) \cup (H_2 \cap E)] \\ &= P(H_1 \cap E) + P(H_2 \cap E). \\ &= P(H_1) \cdot P(E/H_1) + P(H_2) \cdot P(E/H_2) \end{aligned}$$

This is a particular case of the law of total probability. This result can be extended and we have the law of total probability.

Law of Total Probability

If $H_1, H_2, \ldots, H_n$ are mutually exclusive and exhaustive events (Fig. 11.5) and E is any event that occurs with H_1 or $H_2, \ldots, H_n$ then

$$P(E) = P(H_1)\,P(E/H_1) + P(H_2)\,P(E/H_2) + \ldots + P(H_n)\,P(E/H_n)$$

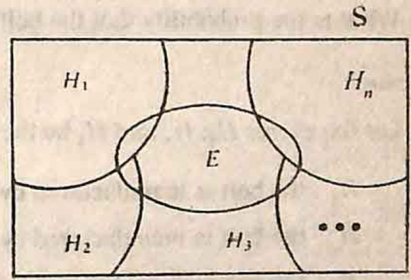


Fig. 11.5

We know that
$P(H_1 \cap E) = P(H_1) \cdot P(E/H_1)$
Interchanging H_1 and E, we have:
$P(E \cap H_1) = P(E) \cdot P(H_1/E)$
But $H_1 \cap E = E \cap H_1$
Therefore,

$$P(E) \cdot P(H_1/E) = P(H_1) \cdot P(E/H_1)$$

$$\Rightarrow \quad P(H_1/E) = \frac{P(H_1) \cdot P(E/H_1)}{P(E)}$$

$$= \frac{P(H_1) \cdot P(E/H_1)}{P(H_1)\,P(E/H_1) + P(H_2)\,P(E/H_2)}$$

This result is a particular case of Bayes' Formula which can be extended.

Bayes' Formula

If $H_1, H_2, \ldots, H_n$ *are mutually exclusive and exhaustive events and* E *is any event that occurs with* $H_1, H_2, \ldots, H_n$ *then*

$$P(H_i/E) = \frac{P(H_i) \cdot P(E/H_i)}{P(H_1) \cdot P(E/H_1) + P(H_2) \cdot P(E/H_2) + \ldots + P(H_n)\,P(E/H_n)}$$

Note: You may see that in the *LHS* of the Bayes' formula you have H_i/E and in the *RHS*, you have E/H_i for $i = 1$ to n.

Example 11.16

In a bolt factory, machines, A, B and C manufacture respectively 25%, 35% and 40% of the total bolts. Of their output 5, 4, and 2 per cent are respectively defective bolts. A bolt is drawn at random from the product.

(i) If the bolt drawn is found to be defective, what is the probability that it is manufactured by the machine B ?

(ii) What is the probability that the bolt drawn is defective ?

Solution

(i) Let the events H_1, H_2 and H_3 be the following:

H_1 : the bolt is manufactured by machine A

H_2 : the bolt is manufactured by machine B

H_3 : the bolt is manufactured by machine C.

Clearly H_1, H_2, H_3 are mutually exclusive and exhaustive.

Let the event E be:

E : the bolt is defective

The event E occurs with H_1, with H_2 and with H_3.

Now $P(H_1)$ = Probability that the bolt drawn is manufactured by machine A

$= 25\% = 0.25$

Similarly,

$$P(H_2) = 0.35 \text{ and } P(H_3) = 0.40$$

Again

$P(E/H_1)$ = Probability that the bolt drawn is defective given the condition that it is manufactured by machine A

$= 5\% = 0.05$

Similarly,

$$P(E/H_2) = 0.04 \text{ and } P(E/H_3) = 0.02$$

We are required to find $P(H_2/E)$, i.e. given the condition that the bolt drawn is defective, what is the probability that it was manufactured by machine B.

We have the Bayes' formula

$$P(H_2/E)=\frac{P(H_2)P(E/H_2)}{P(H_1)\cdot P(E/H_1)+P(H_2)\cdot P(E/H_2)+P(H_3)\cdot P(E/H_3)}$$

Substituting in the above expression

$$P(H_2/E)=\frac{0.35\times 0.04}{0.25\times 0.05+0.35\times 0.04+0.40\times 0.02}=\frac{0.0140}{0.0345}=\frac{28}{69}$$

(ii) Probability that the bolt drawn is defective $= P(E)$

$$= P(H_1)\cdot P(E/H_1)+P(H_2)\cdot P(E/H_2)+P(H_3)\cdot P(E/H_3)$$

$$= 0.25\times 0.05+0.35\times 0.04+0.40\times 0.02$$

$$= 0.0345$$

EXERCISE 11.10

1. A bag X contains 2 white and 3 red balls and a bag Y contains 4 white and 5 red balls. One ball is drawn at random from one of the bags and is found to be red. Find the probability that it was drawn from bag Y.

2. Three urns contain 6 red, 4 black; 4 red, 6 black, and 5 red and 5 black balls respectively. One of the urn is selected at random and a ball is drawn from it. If the ball drawn is red, find the probability that it is drawn from the first urn.

3. The contents of Urns I, II, III are as follows:

   Urn I : 1 white, 2 black and 3 red balls.
   Urn II : 2 white, 1 black and 1 red balls.
   Urn III : 4 white, 5 black and 3 red balls.

   One urn is chosen at random and two balls are drawn. They happen to be white and red. What is the probability that they come from Urns I, II, III ?

4. A factory has two machines A and B. Past records show that the machine A produced 60% of the items of output and machine B produced 40% of the items. Further 2% of the items produced by machine A were defective and 1% produced by machine B were defective. If a defective item is drawn at random, what is the probability that it was produced by the machine A ?

5. A company has two plants to manufacture scooters. Plant I manufactures 70% of scooters and Plant II manufactures 30%. At Plant I, 80% of the scooters are rated as of standard quality and at Plant II, 90% of the scooters are rated as of standard quality.

A scooter is chosen at random and is found to be of standard quality. What is the probability that it has come from Plant II ?

6. An Insurance Company insured 2000 scooter drivers, 4000 car drivers and 6000 truck drivers. The probability of an accident involving a scooter, car and a truck is 0.01, 0.03 and 0.15 respectively. One of the insured persons meets with an accident. What is the probability that he is a scooter driver ?

C. RANDOM VARIABLES

11.14 Random Variables

A random variable is often described as a variable whose values are determined by chance. We shall now try to make this idea precise.

The word "variable" indicates that we are concerned with something which cannot be given a fixed value but can assume different values under different situations. The expression "determined by chance" means that the values of the variables are determined by the outcomes of a random experiment, that is the value taken by the variable is known only when the outcome of a random experiment is known. The following example will make the idea clearer.

Suppose we have 10 families $F_1, F_2, \ldots, F_{10}$ having 3, 4, 3, 2, 5, 4, 3, 6, 4, 5 members respectively. Let our random experiment consist in selecting one family out of the ten in such a manner that each family is equally likely to be selected. Then, the probability that a particular family has been selected is the same for all families and is equal to $\frac{1}{10}$.

Let us select a family and then count the number of members in the family, denoting the number by X. If F_1 is the family selected we will have $X = 3$, if F_2 is selected $X = 4$, and so on. We say that X can take any of the values 2, 3, 4, 5 and 6, the value actually taken being known only after the selected family is known. Thus, X is a *variable* because it has different possible values; it is a random variable because the value it actually takes is known only after the random experiment of selecting a family has been carried out and its outcomes observed.

If you recall the general definition of a *function* you will realise that the random variable X is, really speaking, a function whose *domain* is the set of outcomes $\{F_1, F_2, \ldots, F_{10}\}$ of a random experiment with

$$X(F_1) = 3, X(F_2) = 4, X(F_3) = 3, X(F_4) = 2, X(F_5) = 5,$$
$$X(F_6) = 4, X(F_7) = 3, X(F_8) = 6, X(F_9) = 4, X(F_{10}) = 5.$$

We may also say that the *co-domain* of the function X is the set of natural numbers, or the set of integers, or even the set of real numbers, and that its *range* is the set $\{2, 3, 4, 5, 6\}$.

Consider another random experiment which consists in tossing 3 coins. The sample space of the eight possible outcomes of this experiment may be described by the set

$$S = \{HHH, HHT, HTH, HTT, \\ THH, THT, TTH, TTT\}.$$

Having performed the experiment let us count the number of heads observed and denote this number by X. Then we have $X = 0$ if the outcome of the experiment is TTT, $X = 1$ if the outcome is HTT, THT or TTH, and so on. The possible values of X are 0, 1, 2 and 3. We see that X is a random variable whose values are determined by the results of the random experiment of tossing three coins. We can also regard X as a function with domain **S**, and range $\{0, 1, 2, 3\}$. The values of X for different outcomes of the experiment are given by

$$X(HHH) = 3, X(HHT) = 2, X(HTH) = 2, X(HTT) = 1, \\ X(THH) = 2, X(THT) = 1, X(TTH) = 1, X(TTT) = 0.$$

We are now in a position to give the general definition of a random variable, which is as follows:

Definition

A random variable is a real valued function whose domain is the sample space of a random experiment.

You will recall that two functions are regarded as being different if either the domains or the ranges are different, or, if the domains or ranges are the same, the values of the two functions are different at least at one point of the common domain. Since random variables are also functions we will regard two random variables X and Y as different if

(i) the sample spaces of the associated random experiments are different, or
(ii) the sets of possible values of the two random variables are different, or
(iii) if the associated sample spaces and sets of possible values are the same, X and Y take different values at least at one point of the common sample space.

Example 11.17

A coin is tossed twice. X denotes the number of heads and Y the number of tails. The two random variables have the same domain $\{HH, HT, TH, TT\}$ and the same range {0, 1, 2) but are different since

$$X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0$$

and

$$Y(HH) = 0, Y(HT) = 1, Y(TH) = 1, Y(TT) = 2.$$

Example 11.18 ·

A and B play a game in which each throws a die. If the number on A's die is even he gets Rs 2 from B, if it is odd but greater than the number on B's die A gets Re 1 from B. In all other cases A pays Rs 3 to B. If X denotes the amount won by A (a negative value of X indicating a loss), then X is a random variable with range $\{-3, 1, 2\}$. The domain of X consists of 36 elements of the sample space and X is the function described by

$X(1, 1) = -3, X(1, 2) = -3, \ldots, X(1, 6) = -3$
$X(2, 1) = 2, X(2, 2) = 2, \ldots, X(2, 6) = 2$
$X(3, 1) = 1, X(3, 2) = 1, X(3, 3) = -3, \ldots, X(3, 6) = -3$
$X(4, 1) = 2, X(4, 2) = 2, \ldots, X(4, 6) = 2$
$X(5, 1) = 1, X(5, 2) = 1, X(5, 3) = 1, X(5, 4) = 1, X(5, 5) = -3, X(5, 6) = -3$
$X(6, 1) = 2, X(6, 2) = 2, \ldots, X(6, 6) = 2.$

EXERCISE 11.11

1. A bag contains 3 red and 4 black balls. One ball is drawn and then put back in the bag. The process is repeated three times. Every time the ball drawn happens to be red we say that the draw has resulted in a "success". Let X denote the number of successes recorded in 3 draws. Show that X can be considered as a random variable and exhibit it as a function on the sample space of the experiment.
2. If in Exercise 1 we assume that the ball drawn is not put back in the bag, then what is the representation of X as a function on the sample space ?

11.15 Probability Distribution of a Random Variable

Let us look again at the example of 10 families $F_1, F_2, \ldots, F_{10}$ with 3, 4, 3, 2, 5, 4, 3, 6, 4, 5 members respectively. The random variable X in this case could take any of the values 2, 3, 4, 5 and 6, depending on which family is selected. Out of the 10 families, there is only one family (F_4) with 2 children, three families (F_1, F_3, F_7) have three children each, three (F_2, F_6, F_9) have four children, two (F_5, F_{10}) have five, and only one (F_8) has six children. So, X will get the value 2 only when family F_4 is selected. Since we had assumed that each family is equally likely to be selected, the probability that F_4 is selected is $\frac{1}{10}$. We can thus also say that the probability that X takes the value 2 is equal to $\frac{1}{10}$. Again, the random variable X will have the value 3 when the selected family is either F_1 or F_3 or F_7. Thus, we can say that the probability that X takes the value 3 is equal to the probability that the family selected is F_1, F_3, or F_7, that is, it is equal to $\frac{3}{10}$. Continuing in this manner we obtain

$$P[X = 2] = \frac{1}{10}, \quad P[X = 3] = \frac{3}{10}, \quad P[X = 4] = \frac{3}{10}$$

$$P[X = 5] = \frac{2}{10}, \quad P[X = 6] = \frac{1}{10}.$$

Such a description giving the values of the random variables along with the corresponding probabilities is called the *probability distribution of the random variable*.

Example 11.19

Let us call X the random variable which equals the number of heads obtained when 3 coins are tossed. There are eight possible outcomes of tossing 3 coins and we may consider the eight outcomes as equally likely so that every single outcome has probability equal to $\frac{1}{8}$.

Obviously, the possible values of X are 0, 1, 2 and 3. X takes the value 0 when the outcome is TTT, the value 1 when the outcome is HTT, THT or TTH, and so on. So the probability of X taking the value 0 may be taken to be equal to the probability that the outcome of the random experiment is TTT. Arguing in this manner we get

$$P(X=0) = P(TTT) = \frac{1}{8}$$

$$P(X=1) = P(HTT, THT \text{ or } TTH) = \frac{3}{8}$$

$$P(X=2) = P(HHT, HTH \text{ or } THH) = \frac{3}{8}$$

$$P(X=3) = P(HHH) = \frac{1}{8},$$

which describes the probability distribution of X.

The probability distribution of a random variable may be defined as follows:

Definition

The probability distribution of a random variable X is the system of numbers

$$\begin{pmatrix} x_1, x_2, \ldots, x_n \\ p_1, p_2, \ldots, p_n \end{pmatrix}, p_i > 0, \quad \sum_{i=1}^{n} p_i = 1$$

where the real numbers $x_1, x_2, \ldots, x_n$ are the possible values of the random variable X and p_i ($i = 1, 2, \ldots, n$) is the probability of the random variable X taking the value x_i.

The probability distributions of the random variables described in the two examples earlier may therefore be written as

$$\begin{pmatrix} 2 & 3 & 4 & 5 & 6 \\ \frac{1}{10} & \frac{3}{10} & \frac{3}{10} & \frac{2}{10} & \frac{1}{10} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$

respectively.

It should be noted that the probability distribution of a random variable can be obtained only when the random experiment, on which the random variable is defined as a function, is known. If x is one of the possible values of a random variable X, the statement $X = x_i$ is true only at some points of the sample space of the random experiment. The subset consisting of these points represents an event E associated with the random experiment. We take the probability of this event E to be the probability that the random variable X takes the value x_i. Going through this process for all the possible values of X we obtain its probability distribution.

Example 11.20

Consider the following random experiment. A pack of playing cards is taken and from it all the cards except Aces, Kings, Queens and Jacks are removed. We are thus left with a pack of 16 cards which is shuffled and then one card is drawn. Each of the 16 cards can be considered to be equally likely to be drawn so that the probability of drawing any given card out of the 16 is equal to $\frac{1}{16}$. The random variable Y is defined as follows. $Y = 0$ if the card drawn is the Jack of Clubs or Diamonds, $Y = 1$ if the card drawn is the Jack of Hearts or Spades or a Queen, $Y = 2$ if the card drawn is a King or the Ace of Clubs or Diamonds, and $Y = 3$ if the card drawn is the Ace of Hearts or Spades. Then the probability that Y takes the value 0 is equal to the probability that the card drawn is the Jack of Clubs or Diamonds. Therefore,

$$P(Y = 0) = \frac{2}{16} = \frac{1}{8}.$$

Proceeding in this manner for the other values of Y we find that its probability distribution is

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$

You may have noticed that the probability distribution of the random variable Y is the same as that of the random variable X defined as the number of heads obtained on tossing 3 coins. We thus see that two different random variables can have the same probability distribution.

We could have constructed an example of a random variable having the same probability distribution as that of X in a simpler manner as follows.

Example 11.21

We take eight identical slips of paper, write the number 0 on one of them, the number 1 on three of the slips, the number 2 on three of the slips, and the number 3 on one of the slips. The eight slips of paper are then folded together (as one does for a lottery), put in a box and thoroughly mixed. One slip is then picked up from the box. Each one of the eight slips is equally likely to be selected. Let us take as our random variable Z the number written on the selected slip. One verifies easily that the probability distribution of Z is the same as that of X namely,

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$

The example of the random variable Z also illustrates the important fact that if we know only the probability distribution of a random variable we can also define a random experiment and a random variable as a function on its sample space whose probability distribution is the same as that of the given one.

It should also be noted that two random variables with different sets of possible values cannot have the same probability distribution. In the language of functions, if the ranges are different the random variables cannot have the same probability distribution; but, if the ranges are the same the random variables can have the same probability distribution even though their domains may be different.

Can you construct an example of two random variables which are defined as functions on the same sample space, have the same set of possible values, but have different probability distributions ?

EXERCISE 11.12

1. A class has 15 students whose ages are 14, 17, 15, 14, 21, 19, 20, 16, 18, 17, 20, 17, 16, 19, and 20 years respectively. One students is selected in such a manner that each has the same chance of being selected and the age X of the selected student is recorded. What is the probability distribution of the random variable X ?

2. In Exercise 1 find the probability that
   (i) the age of the selected student is divisible by 3.
   (ii) the age of the selected student is more than 16.
   (iii) the selected student is eligible to vote in the General Election.

11.16 Binomial Distribution

Consider a random experiment with sample space S and an event E (a subset of S) associated with it. Then the event "not-E " may be denoted by E^c, the complement (in S) of the subset E. Let $P(E) = p$, $P(E^c) = q$; so that $p, q > 0$ and $p + q = 1$.

If the experiment results in the event E we say a "success", denoted by S, has occurred. If on the other hand, the event E does not occur (i.e. the complementary event E^c occurs) we say the experiment has resulted in a "failure", denoted by F. We can then say that the probability of a success is equal to p and that of a failure is $1 - p$.

Suppose the experiment is carried out twice under identical conditions so that we can regard them as independent experiments. If we consider the two experiments as a single experiment there are the following four possibilities of the occurrence of E and E^c:

(i) E occurs in the first, E occurs in the second
(ii) E occurs in the first, E^c occurs in the second
(iii) E^c occurs in the first, E occurs in the second
(iv) E^c occurs in the first, E^c occurs in the second.

In terms of "success" and "failure" the four outcomes of the combined experiment can be written as

$$SS, SF, FS, FF.$$

Since, the two experiments are independent, the probabilities of these four outcomes are given by

$$P(SS) = P(S) \cdot P(S) = p^2$$
$$P(SF) = P(S) \cdot P(F) = pq$$
$$P(FS) = P(F) \cdot P(S) = qp$$
$$P(FF) = P(F) \cdot P(F) = q^2.$$

Note that the sum of these four probabilities

$$p^2 + pq + qp + q^2 = 1.$$

Can you see why it is so ?

Let us now define a random variable X, on the sample space of the four outcomes given above, as the "number of successes". The possible values that the random variable takes are 0, 1 and 2. The probability distribution of X is given by

$$P(X = 0) = P(FF) = q^2$$
$$P(X = 1) = P(SF \text{ or } FS) = 2pq$$
$$P(X = 2) = P(SS) = p^2$$

If the experiment is carried out three times under identical conditions, then for the combined experiment there are eight possible outcomes. In terms of success (S) and failure (F) these eight outcomes are given by

$$SSS, SSF, SFS, SFF$$
$$FSS, FSF, FFS, FFF.$$

Since the experiments are independent, the probabilities of these outcomes are

$$P(SSS) = p^3, \qquad P(SSF) = p^2q$$
$$P(SFS) = pqp = p^2q, \qquad P(SFF) = pq^2$$
$$P(FSS) = qp^2 = p^2q, \qquad P(FSF) = qpq = pq^2$$
$$P(FFS) = q^2p = pq^2, \qquad P(FFF) = q^3.$$

The sum of these probabilities is

$$p^3 + 3p^2q + 3pq^2 + q^3 = (p + q)^3 = 1.$$

Let X now stand for the random variable "number of successes" defined on the sample space of this experiment. Then

$$\begin{aligned} P(X=0) &= P(FFF) = q^3 \\ P(X=1) &= P(SFF, FSF, FFS) = 3pq^2 \\ P(X=2) &= P(SSF, SFS, FSS) = 3p^2q \\ P(X=3) &= P(SSS) = p^3 \end{aligned}$$

describes the probability distribution of X. Note that in the two examples considered above the probability values in the probability distribution of X are equal to the different terms of the binomial expansions of $(q+p)^2$ and $(q+p)^3$ respectively.

The above results can be easily generalised to the case where the experiment is repeated n times under identical conditions. If the success (occurrence of E) and failure (non-occurrence of E) are recorded successively as the experiment is repeated, we will get a result of the type

$$S\,F\,F\,S\,S\,F\ldots F\,S$$

There are 2^n such outcomes which constitute the sample space of the combined experiment. Since the experiments are independent, the probability of the outcome above is

$$p \cdot q \cdot q \cdot p \cdot p \cdot q \ldots q \cdot p = p^r q^{n-r}$$

if r is the number of successes in the outcome. As before, we define X to be the random variable "number of successes". Then the possible values of X are $0, 1, 2, \ldots, n$.

Let us now try to calculate the probability that X takes the value r. We have

$$P(X=r) = P(\text{the outcome of the experiment consists of } r \text{ successes and } (n-r) \text{ failures}).$$

Now each of the 2^n outcomes of the combined experiment which has r successes and $(n-r)$ failures has probability $p^r q^{n-r}$. Also, the total number of outcomes with exactly r successes is the same as the number of ways in which r out of n positions can be selected, that is, it is equal to $C(n, r)$. Thus we have

$$P(X=r) = C(n,r)\,p^r q^{n-r}$$

The probability distribution of the random variable is therefore given by

$$\begin{pmatrix} 0 & 1 & 2 & \ldots & r & \ldots & n \\ q^n & C(n,1)\,pq^{n-1} & C(n,2)\,p^2q^{n-2} & \ldots & C(n,r)\,p^r q^{n-r} & \ldots & p^n \end{pmatrix}$$

We see that the probabilities of the random variable taking values $0, 1, 2, \ldots, n$ are given by the terms in the binomial expansion of $(q+p)^n$. Because of it we say that the probability distribution of the random variable X is the *Binomial Distribution* or, that X is a *Binomial Random Variable*. We may define the Binomial Distribution as follows:

Definition

The Binomial Distribution $B(n, p)$ is the probability distribution of a random variable which takes values $0, 1, 2, \ldots, n$ with probabilities

$$C(n, 0)p^0q^n,\ C(n, 1)p^1 q^{n-1},\ C(n, 2)p^2q^{n-2}, \ldots, C(n, r)p^r q^{n-r}, \ldots, C(n, n)p^n q^0,$$

where $p, q > 0$ and $p + q = 1$.

Example 11.22

Suppose a coin is tossed 5 times and the random variable X is the number of heads observed. Regarding the appearance of heads in any toss as a success, we see that the probability distribution of X is the Binomial Distribution $B\left(5, \frac{1}{2}\right)$ so that

$$P[X = 0] = C(5, 0)\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{5-0} = \frac{1}{32}$$

$$P[X = 1] = C(5, 1)\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{5-1} = \frac{5}{32}$$

$$P[X = 2] = C(5, 2)\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{5-2} = \frac{10}{32}$$

and so on.

Example 11.23

Consider a group of 10 families each having 5 children of which 2 are boys and 3 are girls. One child is selected from each family, each child of the family being equally likely to be selected. We wish to calculate the probability that among the 10 children so selected there are exactly 5 girls. We may consider this as the case of 10 independent experiments in each of which the probability of success (selected child is a girl) is $\frac{3}{5}$. The number of successes is thus a Binomial random variable with distribution $B\left(10, \frac{3}{5}\right)$ Hence the required probability is

$$C(10, 5)\left(\frac{3}{5}\right)^5\left(\frac{2}{5}\right)^5 = 0.2 \text{ (approx.)}.$$

EXERCISE 11.13

1. A die is thrown 6 times. If "getting an odd number" is a "success", what is the probability of
   (i) 5 successes ?

(ii) at least 5 successes ?
(iii) at most 5 successes ?

2. A pair of dice is thrown 4 times. If getting a doublet is considered a success, find the probability of 2 successes.

3. There are 5 per cent defective items in a large bulk of items. What is the probability that a sample of 10 items will include not more than one defective item ?

4. The items produced by a firm are supposed to contain 5% defective items. What is the probability that a sample of 8 items will contain less than 2 defective iems ?

5. Five cards are drawn successively with replacement from a well-shuffled deck of 52 cards. What is the probability that
   (i) all the five cards are spades ?
   (ii) only 3 cards are spades ?
   (iii) none is a spade ?

6. The probability that a bulb produced by a factory will fuse after 150 days of use is 0.05. Find the probability that out of 5 such bulbs
   (i) none
   (ii) not more than one
   (iii) more than one
   (iv) at least one
   will fuse after 150 days of use.

7. A bag contains 10 balls each marked with one of the digits 0 to 9. If four balls are drawn successively with replacement from the bag, what is the probability that none is marked with the digit 0 ?

MISCELLANEOUS EXERCISE ON CHAPTER 11

1. (a) Two dice are thrown. Describe the sample space of this experiment.
   (b) If E is the event: "sum of numbers appearing on the two dice is even", and F the event: "at least one die shows the number 2", describe the sets representing the events E, F, E or F, only E.
   (c) Assuming all the elementary events to be equally likely, obtain the probabilities of the events: "at least one of the events E and F occurs", "none of the events E and F occurs", "only E occurs", "both E and F occur"

2. E, F and G are three events associated with the sample space S of a random experiment. If E, F and G also denote the subsets of S representing these events, what are the sets representing the events
   (a) Out of the three events at least two events occur

18. A bag contains 5 white, 7 red and 8 black balls. If four balls are drawn one by one with replacement, what is the probability that
    (i) none is white ?
    (ii) all are white ?
    (iii) only 2 are white ?

19. A box contains 100 tickets each bearing one of the numbers from 1 to 100. If 5 tickets are drawn successively with replacement from the box, find the probability that all the tickets bear numbers divisible by 10.

20. An urn contains 25 balls of which 10 balls bear a mark 'X' and the remaining 15 bear a mark 'Y'. A ball is drawn at random from the urn, its mark noted down and it is replaced. If 6 balls are drawn in this way, find the probability that
    (i) all will bear 'X' mark
    (ii) not more than 2 will bear 'Y' mark
    (iii) the number of balls with 'X' mark and 'Y' mark will be equal
    (iv) at least one ball will bear 'Y' mark.

21. The letters of the word 'SOCIETY' are placed at random in a row. What is the probability that the three vowels come together ?

22. Three groups of children contain 3 girls and 1 boy; 2 girls and 2 boys; 1 girl and 2 boys. One child is selected at random from each group. Show that the probability that the three selected consists of 1 girl and 2 boys is $\frac{3}{8}$

23. A can hit a target 3 times in 6 shots, B: 2 times in 6 shots and C: 4 times in 4 shots. They fix a volley. What is the probability that at least 2 shots hit ?

CHAPTER 12

Correlation and Regression

A. STATISTICAL RELATIONSHIP BETWEEN VARIABLES

12.1 Bivariate Frequency Distributions

In the chapter on frequency distributions (Chapter 17 of the textbook for Class XI) we had introduced the terms "*unit of observation*" and " *variable of observation*". It was explained there that each *recorded*, or *observed*, value of a variable of observation is associated with a particular person, place, object, etc, and that the term unit of observation is used to describe what the values of a variable of observation are attached to. Thus, when we talk of data concerning examination results the units of observation are students, and the variable of observation is "total marks obtained"; each recorded value of the variable denotes the total marks obtained by a particular student who has appeared in that examination. Thus distinction between variable of observation and units of observation was very useful in describing the form and content of a frequency table.

In the different examples of frequency tables, as well as in the discussion about frequency tables in Chapter 17 of the textbook for Class XI, we had only one variable of observation in each case. However, while describing the information provided by the census table we had pointed out that to each unit of observation, that is, a person alive at the time of the census, not one but three variables were associated which were age, sex, and place of residence. So, we can have data in which more than one variable of observation may be associated with each unit of observation. It must be clearly understood that the same variables are being considered for *all* the units of observation. For example, if for some students we are recording the age, and for some of them we are recording the height, then we cannot say that we have two variables associated with *each* unit of observation. Such data will have to be studied in two parts. In one we will only consider the group of students with their ages, and in the other we take the group of students with their heights.

When we have more than one variable of observation for which values are being observed for each unit of observation, we say that we have *multivariate data*. In this chapter we shall confine ourselves to the study of those situations where there are only two variables of observation. For such cases we will use the term *bivariate data*. We could have used the term *univariate data* for those cases in which there was only one variable of observation, but the term univariate is usually omitted.

As an example of bivariate data let us go back to the census table (Table 17.1 of Class XI textbook), taking into consideration only two variables, age and sex. The raw data would consist of a record of the age and sex of each person and would look somewhat like the following:

$$(34, M), (11, M), (61, F), \ldots$$

For each individual we show the recorded value of the variables as an ordered pair. The raw data presented in the form above states that one individual was a male aged 34 years, another was a male aged 11 years, the next was a female aged 61 years, and so on.

In general, if the two variables are denoted by X and Y, with possible values $x_1,\ldots, x_m$, and $y_1,\ldots, y_n$ respectively, the raw data would be represented by ordered pairs like (x_1, y_1), (x_3, y_2), (x_1, y_2), and so on, there being one ordered pair of values of the variables for each unit of observation. There are mn possible ordered pairs of values (x_i, y_j), though not every one of these mn pairs may be observed in any particular situation.

The bivariate raw data is usually presented in the form of a table called "*the bivariate frequency table*". The word *bivariate* indicates that we have *two* variables of observation; the words "*frequency table*" indicate that we shall be using a form of presentation of data similar to that used in frequency table in Chapter 17 of Class XI textbook, which would be more accurately described as *univariate* frequency tables. Just as a (univariate) frequency table had also been given the name "*frequency distribution*", in the same way we can use the words " bivariate frequency distribution" in place of "bivariate frequency table". As in the univariate case the use of the word distribution indicates the process of what is being done, and the use of the word table refers to what the result of following that process looks like .

As example of a bivariate frequency table let us take a look at the census table (Table 17.1 of Class XI textbook), taking into consideration only the age and sex variable. With these two variables the table will look somewhat as follows.

TABLE 12.1

Age-group	*Male*	*Female*	*Total*
All ages	47,016,421	41,324,723	88,341,144
0 – 9	13,724,165	12,381,238	26,105,403
10 – 14	6,061,626	4,798,307	10,859,993
15 – 19	3,978,135	3,206,413	7,184,548
...	...	...	...
...	...	...	...
70 +	1,145,516	953,493	2,099,009
Age not stated	3,723	3,538	7,261

Let us now recall what was said earlier in Chapter 17 of Class XI textbook about the method of constructing a frequency table from raw data. We use the *values* of the variable of observation to define a number of classes and then *count* the number of units of observation for which the value of the variable fall in a given class; the number of units so obtained is the frequency of that class.

If we examine closely the table above we will notice that something similar is happening here. The only difference is that our classes are now being defined by the values of *two* variables instead of one. Thus, one class is defined by the values 0 – 9 of the age variable and the value "male" of the sex variable. The frequency of this class is 13,724,165 meaning thereby that there were 13,724,165 persons (units of observation) who were males *and* had ages from 0 to 9 years. Similarly, there were 3,206,413 persons who were females *and* had ages from 15 to 19 years. We see that the values of both the variables are being used to define the classes of the frequency table.

In the above table the rows correspond to classes defined by the values of one of the variables (the age variable), and the columns give the classes defined by the values of the other variable (the sex variable). This is the usual form in which a bivariate frequency table is presented. The form of presentation can sometimes be different but the classes will *always* be defined by using the values of both the variables.

The first row ("All ages") gives the totals of the frequencies in the columns of the table. In most bivariate frequency tables these totals are written in the last row. The last column ("Total") gives the totals of the frequencies in the rows of the table. By examining these two totals we find that there were 88,341,144 persons (units of observation) in all, of which 47,016,421 were males and 41,324,723 were females. Similarly, there were in all 7,184,548 units of observation whose ages were between 15 and 19 years and out of them 3,978,135 were males, and 3,206,413 were females. The raw data when put in the form of a bivariate frequency table not only exhibits the information in a compact form but also highlights the important points. For example, a few simple calculations show that whereas the males constitute 55.4% (47, 016,421 ÷ 88,341,144) of the whole population, they form only 52.6% (13,724,165 ÷ 26,105,403) of the 0 to 9 age-group. The percentage of males increases to 54.6 in the 70+ age-group. In the same way, we find that 29.2% (13,724,165 ÷ 47,016,421) of all the males are in the age-group 0 to 9, and 30% (12,381,238 ÷ 41,324,723) of all the females are in the 0 to 9 age-group. When we come to the 10 to 14 age-group we find that 12.9% of all the males, and only 11.6% of all the females are in this age-group.

In the bivariate frequency table considered by us, one of the variables (the age variable) was a *quantitative* variable, and the other variable (the sex-variable) was a *qualitative* variable. We can also have bivariate frequency tables in which both variables are qualitative, or both are quantitative variables.

Example 12.1

The following is an example of a bivariate frequency table in which both variables are qualitative.

TABLE 12.2

WORKERS AND NON-WORKERS (1981) SEX-WISE – RURAL INDIA (NO. IN MILLIONS)

Sex	*Main Workers*	*Marginal Workers*	*Non-workers*	*Total*
Male	134.1	5.4	117.3	256.8
Female	40.4	18.1	186.6	245.1
TOTAL	174.5	23.5	303.9	501.9

Source: *Census of India*, 1981; Paper 3, Provisional Population Tables
(Taken from Table 3.3, *Rural Development Statistics*, NIRD, Hyderabad, 1985, p. 38)

One of the variables in this table is the sex variable with values "male" and "female". The other variable may be called "work status" with three values "main workers", "marginal workers", and "non-workers". Using the values of these two variables we have six classes with class frequencies (in millions) equal to 134.1, 5.4, 117.3, 40.4, 18.1 and 186.6. The total number of units of observation (i.e., persons living in rural areas) is 501.9 million of which 51.2% are males and 48.8% are females. Among the males 45.7% (117.3 ÷ 256.8) are non-workers, whereas among the females the percentage of non-workers is as high as 76.1% (186.6 ÷ 245.1). . Among the main workers 76.8% (134.1 ÷ 174.5) are males and only 23.1% (40.4 ÷ 174.5) are females. But females constitute the larger share among the marginal workers and non-workers, the percentage of women in these categories being 77% (18.1 ÷ 23.5) and 61.4% (186.6 ÷ 303.9) respectively.

As in the univariate case, care has to be taken while defining the classes of a bivariate frequency table to ensure that no possible pair of values belongs to more than one class, and that every possible pair of values belongs to some class or the other. The simplest way to ensure this is to first define classes separately for each of the two variables and then from them construct classes for pairs of values. We illustrate this by an example in which both variables are quantitative.

Example 12.2

The raw data for this example consists of measurements (in cm) on the sitting height X, and the height Y, for 792 individuals. The sitting height is the height measured from the seat of the chair, on which the person is sitting, to the top of the head. We divide the values of X into 3 classes and those of Y into 5 classes. The class 80– for X represents values of X equal to 80 cm or more but less than 83 cm, whereas the class 160– for Y represents all values of Y equal to 160 cm or more but less than 165 cm. While dividing the values of X and Y into classes we make sure, as in the univariate case, that every possible value of X, or Y is included in some class and that no value belongs to more than one class. This division of values of X into three classes, and of Y into 5 classes, gives us 15 classes in all for the bivariate frequency table, and the total frequency 792 will be distributed in these 15 classes. The bivariate frequency table obtained from the raw data is shown below.

TABLE 12.3

FREQUENCY DISTRIBUTION OF PERSONS BY HEIGHT AND SITTING HEIGHT
(BOTH MEASURED IN CENTIMETRES)

X \ Y	150–	155–	160–	165–	170–	TOTAL
77–	18	23	4	1	1	47
80–	27	59	33	6	3	128
83–	28	194	274	107	14	617
Total	73	276	311	114	18	792

The last column gives the totals of the frequencies in the rows, and the last row gives the totals of the frequencies in the columns. From this table we also see that for the 792 persons the observed values of X were 77 cm or more but less than 86 cm, and those of Y were 150 cm or more but less than 175 cm.

In general, if we have two variables of observation X and Y, and we define m classes using the values of X and n classes using the values of Y, we will get a total of mn classes for the bivariate frequency table. The total frequency (i.e., the total number of units of observation) will then be divided into mn class frequencies of the bivariate frequency distribution.

EXERCISE 12.1

1. What is the number of persons for which, in Table 12.3,
   (a) sitting height is less than 80 cm and height is 155 cm or more but less than 165 cm ?
   (b) sitting height is 80 cm or more and height is less than 160 cm ?

12.2 Marginal Distributions

In the bivariate frequency tables we had also shown the totals of the row frequencies in the last column, and the totals of the column frequencies in the last row. These totals are not an essential part of a bivariate frequency table. Such a table is a description of how the total frequency is divided into the class frequencies of the classes defined by means of the values of the two variables. Hence, it is enough to show the class frequencies and the total frequency to describe a bivariate frequency table.

However, in almost all cases of bivariate frequency tables we also present the totals of the row frequencies in the last column and the totals of the column frequencies in the last row. One reason for doing so is that it provides a check on our calculations; the total of the last column should be equal to the total of the last row as both are equal to the total frequency. For example, in Table 12.3, the total of the last column (47 + 128 + 617) is equal to the total of the last row (73 + 276 + 311 + 114 + 18) and both are equal to the total frequency 792.

But the more important reason for including such totals is that they provided additional information. Let us take a look again at Table 12.3. The last column totals are 47, 128, and 617. The first of these is the number of persons whose sitting height belongs to the class 77–. The second and third totals are respectively the numbers of persons whose sitting heights belong to the classes 80– and 83–. In other words, if we were interested only in the variable X, then the last column would describe the class frequencies of the classes 77–, 80–, and 83–. In the same way, the totals entered in the last row are the class frequencies for the classes 150–, 155–, 160–, 165–, and 170– when we consider only the variable Y.

Thus, the totals entered in the last column, and the last row, give the (univariate) frequency distributions separately for the two variables X and Y. If we had shown these two univariate distributions separately we could get the following tables.

Sitting Height	*Frequency*
77–	47
80–	128
83–	617
Total	792

Height	*Frequency*
150–	73
155–	276
160–	311
165–	114
170–	18
Total	792

In the bivariate frequency table (Table 12.3) we have presented in a compact form these two univariate distributions also, one for the variable X and the other for the variable Y. These two univariate distributions are called the *marginal distributions* of the *bivariate frequency distributions*. The use of the word "marginal" indicates simply that the frequency distributions occupy the margins (last column and last row) of the bivariate frequency tables.

We can now see another advantage in defining the classes of a bivariate frequency table by means of a division of the values of the two variables into classes separately, and in using the row and column presentation. In this way we describe in a single table the bivariate distribution using values of both the variables X and Y as well as the two univariate distributions, one using the values of X only, and the other using the values of Y alone.

12.3 Conditional Distributions

From the bivariate frequency table we can also derive other univariate frequency distributions in addition to the marginal distributions. Consider, for example, the first column of Table 12.3. This column informs us that there were 73 persons whose height was 150 cm or more but less than 155 cm, and out of these 73 persons 18 had sitting height in the class 77–, 27 had sitting height in the class 80–, and 28 had sitting height in the class 83–.

If we are interested only in persons whose height Y is in the class 150-, their frequency distribution by the values of the sitting height X is given by the first column. Similarly, the second column gives the frequency distribution by sitting height X of those persons whose height Y is 155 cm or more but less than 160 cm. Thus each column of the bivariate frequency table itself describe a univariate frequency distribution. We have already discussed the frequency distribution described by the last column and named it the marginal distribution of the values of the variable X. The other three columns also describe frequency distributions with classes defined by the values of the variable X. There is, however, one major difference between these three distributions and the marginal distribution given by the last column. The marginal distribution could have been obtained if we had only taken observations on the variable X. The values of the variable Y are not needed to obtain it. But, for the other three distributions we need the values of the variable Y. Thus, the distribution described by the first column is obtained by the values of X for only those units of observation (73 in number) which have values of Y belonging to the class 150–. For this reason we call the frequency distribution described by the first column as the *conditional distribution* of X given that the values of Y are 150 cm or more but less than 155 cm. The word "conditional" is used to indicate that we are considering only those units of observation for which the values of the variable Y satisfy a certain condition viz. that of being equal to 150 cm or more but less than 155 cm. The first five columns of the bivariate frequency table, excluding the last one, describe five different conditional distributions of X in each of which the units of observation satisfy a different condition regarding the values of the variable Y.

Similarly, each row of the bivariate frequency table, except the last one, gives a frequency distribution for the variable Y in which the units of observation satisfy a certain condition regarding the values of X. For example, the second row describes the frequency distribution by values of Y of those 128 units of observation for which the X values belong to the class 80–. So, the first three rows of the table describe three different conditional distributions of Y in each of which the units of observation satisfy different conditions regarding the values of the variable X.

In general, if the bivariate frequency table has m classes for the X values, and n classes for the Y values, we shall have n conditional distributions for X (one for each Y class), and m conditional distributions for Y (one for each X class).

12.4 Relationship between Variables

You have already come across the idea of relationship between variables through the concept of a function. When we say that the area of a circle is a function of the radius we mean to say that the area and the radius are related variables. If we know the radius we can determine the area, and from the value of the area we can find the value of the radius. Similarly, the height of a body above the ground and the time taken by it to fall freely under gravity to the ground level are related variables: the value of one of them determines the value of the other. We shall call such relationships, where the value of one variable determines that of the other *functional relationships* to distinguish them from another type of relationship between variables discussed below.

Consider the two variables height and weight of a given group of men. We know that in most cases taller people will have greater weight as compared to persons of shorter height. But at the same time, a short fat person may have a greater weight as compared to a lean tall person. Also, we cannot determine the weight of a person from his height though we could try to make a reasonable guess about it. In short, a tall person is more likely to be heavier in weight than a short person though we cannot put this statement as an exact rule to be true in all cases. This is unlike the relationship between radius and area of a circle where we can say that a circle with larger radius will always have a larger area than a circle with a smaller radius. The relationship between height and weight is not a functional relationship of the kind between the radius and area of a circle.

As another example of a relationship between variables which is not a functional relationship consider the data of Table 12.3. We see that persons with high value of the height Y are more likely to have high values for the sitting height X also. But here too no exact rule can be formulated giving the value of X in terms of the value of Y. At the same time we cannot also say that height and sitting height are unrelated.

The relationships which we wish to study are of the type between height and weight, or between height and sitting height. We call them *statistical relationships* to distinguish them from functional relationships. In such relationships we cannot exactly determine the value of one variable from that of the other. But, we can form a general idea of the behaviour of the values of one variable from the given values of the other.

How does one discover if a statistical relationship exists between two variables. In some cases (e.g., height and weight, height and sitting height) we can say so from past experience and knowledge. In other cases the bivariate frequency table gives an indication whether a statistical relationship exists between the variables or not.

Consider the data of Table 12.2. Here we have two variables: sex and work-status. We find that only 16.5% (40.4 ÷ 245.1) of the females are main workers, whereas for the males this percentage is 52.2% (134.1 ÷ 256.8). So a male is more likely to be a main worker than a female. In the same way one can conclude that a female is more likely to be a marginal or non-worker than a male. Looking at the data in another way we find that 76.8% (134.1 ÷ 174.5) of the main workers are males and only 23.2% (40.4 ÷ 174.5) are females. On the other hand the males are 38.6% of the non-workers and the females are 61.4%. Thus, one is more likely to find a male among the main workers, and a female among the non-workers. So, we can say with some justification that the sex and work-status variables are related through a statistical relationship.

We can look at the concept of a statistical relationship between variables in another way. Again suppose we have the data of Table 12.2 before us. If we are told that the person is a main worker and asked to make a guess about the person being male or female we would prefer to answer that the person is a male, but if we are told that the person is a non-worker and again asked to guess whether the person is male or female then our preferred answer would be that the person is a female. We arrive at these answers because we know that a large majority of main workers (76.8%) are males, so that, if from the main worker one worker is selected, the selected worker is more likely to be a male. Similarly, if we select a person out of all persons who are non-workers, the selected person is more likely to be a female.

If we were to summarise the above discussion about statistical relationship between variables, we can say that such a relationship exists if the knowledge of the value of one variable conveys some information about the value of the other variable. Looking again at Table 12.2, if we are not told anything about the work-status of a person, all we can say is that 51.2% (256.8 ÷ 501.9) are males and 48.8% (245.1 ÷ 501.9) are females. But, if we are told that the person belongs to the class of non-workers, then we can say that 38.6% (117.3 ÷ 303.9) are males and 61.4% (186.6 ÷ 303.90) are females. Thus, the information about the value of the work-status variable has conveyed some information about the sex variable in the sense that the relative proportion of males and females is now different from what it was when we had no information about the work status variable.

Special methods have been developed to discover the existence of a statistical relationship between two variables from bivariate data. When both variables in the bivariate data are quantitative we use the term *correlation analysis* to describe the methods designed to find out if a statistical relationship between the two variables exists or not.

If we have reasons to believe that such a relationship exists we are faced with a second problem, which is the estimation of the likely value of one of the variables when the value of the other variable is known. For example, given the height of a person we may wish to estimate the likely weight of the person. In the case of bivariate data with both variables being quantitative variables, the methods of estimating the likely value of one variable from the known value of the other variable form a part of *regression analysis*.

In what follows we shall restrict ourselves to the study of some methods of correlation and regression analysis when both variables are quantitative.

B. CORRELATION ANALYSIS

12.5 Covariance

Before coming to the study of correlation, we introduce the concept of *covariance* between two quantitative variables. This concept will then be used to develop methods of correlation analysis.

Suppose we have n units of observation for each of which we have observations on two quantitative variables X and Y. Let the n pairs of observations be written as the ordered pairs

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

where $x_1, x_2, \ldots$ denote observed values of the variable X, and $y_1, y_2, \ldots$ those of Y. We first obtain the arithmetic means

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} \qquad y = \frac{y_1 + \ldots + y_n}{n}$$

of the observed values of the variables X and Y. Next, we obtain the *deviations*

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$$

of the observed X values from their mean $\bar{x}$, and the deviations

$$y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y}$$

of the observed Y values from their mean $\bar{y}$. We then multiply the two deviations for each unit of observation to get the n products

$$(x_1 - \bar{x})(y_1 - \bar{y}), (x_2 - \bar{x})(y_2 - \bar{y}), \ldots, (x_n - \bar{x})(y_n - \bar{y}).$$

These n products are then added together and divided by n (to get the average value of the n products) to define the covariance of X and Y, which we will write as Cov (X, Y).

Thus we have :

Definition

The covariance Cov (X,Y) between two variables X and Y is given by

$$\text{Cov}(X, Y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

12.6 Calculation of Covariance

Given a set of n pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$ we do not use the above definition to calculate the covariance. We use instead a slightly different formula which makes the calculations easier to carry out and which also reduces the chances of errors. The formula we use for calculating the covariance is based on the algebraic identity

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}$$

$$= \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)$$

A similar method was used for calculating the variance where

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

was replaced by

$$\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2$$

Using the above identity the covariance is calculated as follows:

(i) obtain the sums $\sum x_i, \sum y_i$

(ii) obtain the sum $\sum_{i=1}^{n} x_i\, y_i$ of the products $x_i y_i$

(iii) calculate the difference

$$\sum x_i y_i - \frac{1}{n} \left(\sum x_i \right) \left(\sum y_i \right)$$

and divide by n.

Example 12.3

Let the pairs of observations be

(1, 10), (2, 9), (3, 8), (4, 8), (5, 6),
(6, 12)), (7, 4), (8, 3), (9, 18), (10, 1).

We have

$$\sum x_i = 1 + 2 + \ldots + 10 = 55$$

$$\sum y_i = 10 + 9 + \ldots + 1 = 79$$

$$\sum x_i y_i = 10 + 18 + 24 + 32 + 30 + 72 + 28 + 24 + 162 + 10 = 410$$

$n = 10.$

Hence,

$$\text{Covariance} = \frac{1}{10}\left(410 - \frac{79 \times 55}{10}\right)$$

$$= -2.45$$

Example 12.4

Let the pairs of observations be

(1, 6), (2, 9), (3, 6), (4, 7), (5, 8),
(6, 5), (7, 12), (8, 3), (9, 17), (10, 1).

We have

$$\sum x_i = 1 + 2 + \ldots + 10 = 55$$

$$\sum y_i = 6 + 9 + \ldots + 1 = 74$$

$$\sum x_i y_i = 6 + 18 + 18 + 28 + 40 + 30 + 84 + 24 + 153 + 10 = 411$$

$n = 10.$

Hence,

$$\text{Covariance} = \frac{1}{10}\left(411 - \frac{74 \times 55}{10}\right)$$

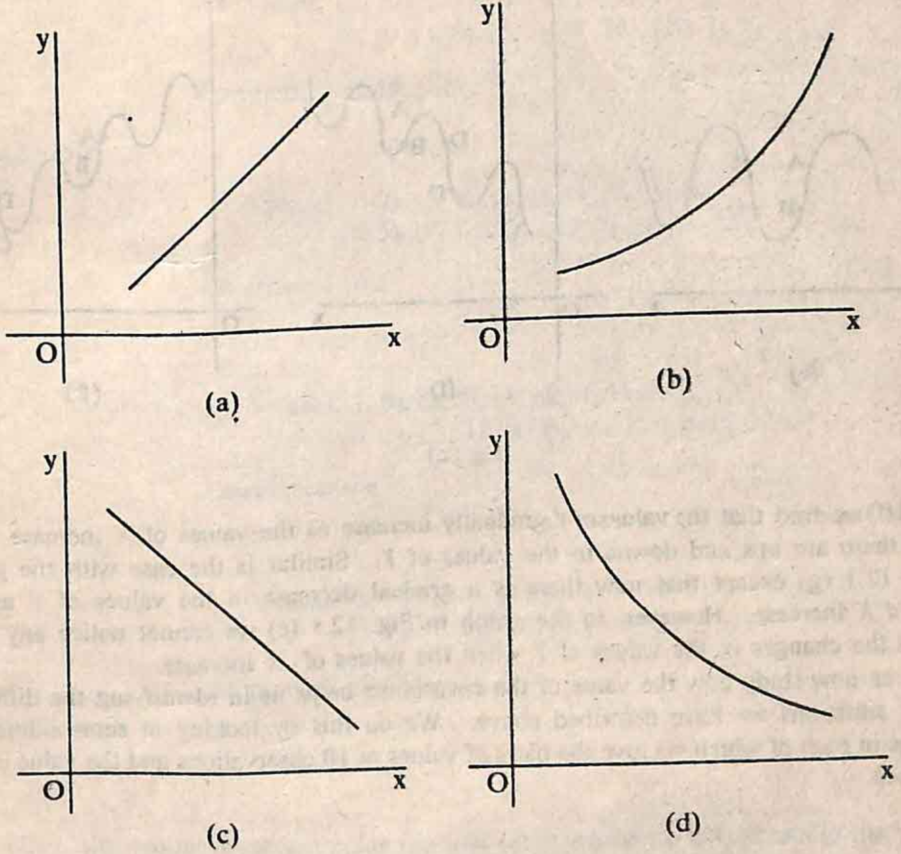
$$= 0.4$$

12.7 Meaning of Covariance

The first thing to notice is that the procedure to obtain the covariance is almost similar to the procedure of obtaining the variance described in Chapter 17 of the text book for Class XI. We recall that to arrive at the definition of variance also we started with the *deviations* $x_1 - x, x_2 - x, \ldots$ of the observed values $x_1, x_2, \ldots$ from their mean x. These deviations were squared, the squared deviations were added together and then divided by the total number of observations to get the average value of the squared deviations which was called the variance of the observed values.

In the present case we have *two* sets of deviations, one for the observed values of X and the other for the observed values of Y. It is these two sets of deviations which are being multiplied together and then averaged. Why is it that we are taking these products ? In the case of the variance we were taking the squares of the deviations because, as we had explained, we were trying to measure the variation in the observed values of the variable, that is, whether the observed values were very much different from each other, or were approximately equal. What are we trying to measure now by working with *products* of the deviations of *two* variables ?

The use of the word "covariance" itself suggests that what we are trying to measure is how the pairs of observations $(x_1, y_1), (x_2, y_2), \ldots$, one pair for each unit of observation, are varying or changing. Is an increase in the values of one variable, say X, accompanied by an increase in the values of the other variable Y ? Or, do the values of one variable decrease when those of the other increase ? Or, do the values of one variable sometimes increase as the values of the other variable increase ? The value of the covariance obtained in any particular case provides a sort of answer to these questions. Let us illustrate these ideas by considering the following examples of *functional relationship* defined by means of graphs of functions.



(a) (b)

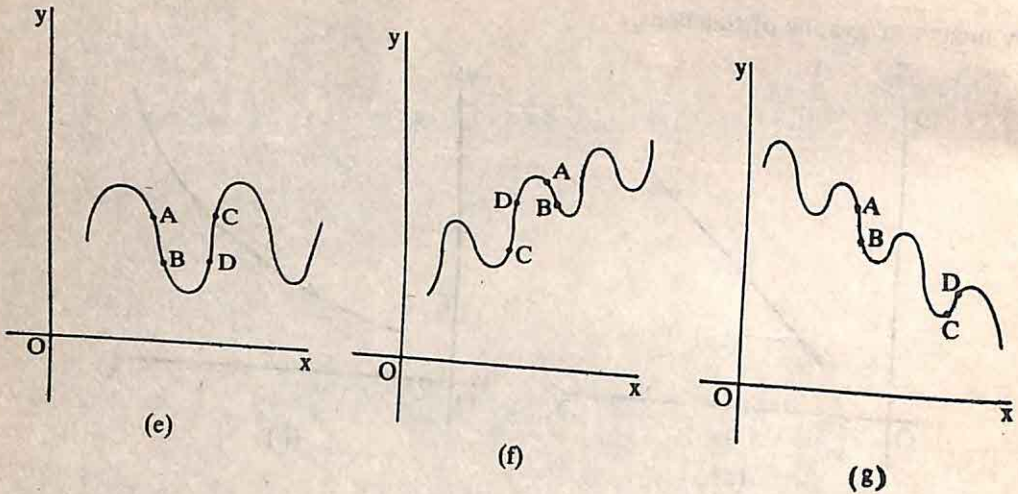
(c) (d)

Fig. 12.1

In Fig. 12.1 (a) we see that the values of X and Y increase or decrease together: if the value of X increases the value of Y also increases, and vice-versa; if the value of X decreases that of Y also decreases, and vice-versa. The same is true for the function described by the graph in Fig. 12.1 (b). The situation described in Figs. 12.1 (c) and (d) is the opposite. Now, the values of one variable decrease if those of the other increase, that is, an increase in the values of X show a decrease in the values of Y, and a decrease in the values of X is accompanied by an increase in the values of Y.

We have a different situation in the graphs given in Figs. 12.1 (e), (f) and (g).

In all these three figures we find that as the values of X increase, the values of Y sometimes decrease (see points A and B on the graphs), and sometimes increase (see points C and D on the graphs). However, there is an important difference when we compare the graphs in Fig. 12.1 (f) and (g) with the graph in Fig. 12.1 (e). In the graph in



(e) (f) (g)

Fig. 12.1

Fig12.1(f) we find that the values of Y gradually increase as the values of X increase even though there are ups and downs in the values of Y. Similar is the case with the graph in Fig. 12.1 (g) except that now there is a gradual decrease in the values of Y as the values of X increase. However, in the graph in Fig. 12.1 (e) we cannot notice any such *trend* in the changes in the values of Y when the values of X increase.

Let us now study how the value of the covariance helps us in identifying the different types of situations we have described above. We do this by looking at some numerical examples in each of which we give the pairs of values of 10 observations and the value of the covariance.

Example 12.5

Values: (1, 2), (2, 4), (3, 6), (4, 8), (5, 10),
(6, 12), (7, 14), (8, 16), (9, 18), (10, 20).
Covariance: 16.5

Example 12.6

Values: (1, 10), (2, 9), (3, 8), (4, 7), (5, 6),
(6, 5), (7, 4), (8, 3), (9, 2), (10, 1).
Covariance: −8.25

Example 12.7

Values: (1, 10), (2, 9), (3, 8), (4, 7), (5, 6),
(6, 5), (7, 4), (8, 3), (9, 18), (10, 1).

Covariance: −2.45

Example 12.8

Values: (1, 2), (2, 4), (3, 8), (4, 7), (5, 10),
(6, 5), (7, 14), (8, 16), (9, 2), (10,20).

Covariance: 10.2

Example 12.9

Values: (1, 6), (2, 9), (3, 6), (4, 7), (5, 8),
(6, 5), (7,12), (8, 3), (9, 17), (10,1).

Covariance: 0.4

Example 12.10

Values: (1, 6), (2, 9), (3, 6), (4, 7), (5, 8),
(6, 5), (7, 11), (8, 3), (9, 15), (10, 1).

Covariance: −0.45

When we look at those examples the first thing we notice is that unlike the variance, which can only take positive values, the values of the covariance can be both positive and negative. Some positive values are large (e.g. 16.5), some are small (e.g. 0.4). Similarly, the negative values can be large in *magnitude* or small (e.g. −8.25 and −0.45).

Let us now look more closely at the example starting with those where the covariance is positive. In example 12.5, an increase in the value of X (the first variable of the pair) is associated with an increase in the value of the second variable Y. In fact, each value of Y is double the corresponding value of X. In example 12.8, we find that as the values of X increase from the first to the tenth pair, the values of Y are mostly increasing, with the fourth, sixth and ninth pairs showing a decrease. The value of covariance is still positive but is lower than that in example 12.5. In the example 12.9, the values Y, as in example 12.8, sometimes increase, and sometimes decrease, as the values of X increase, but as compared to example 12.8 the fluctuations here are larger. The covariance value is now very small.

From these three examples we may conclude, *though without proof*, that large positive values of covariance indicate that the values of the two variables mostly increase or decrease together, though for a few pairs of values this may not be true. As the positive value of the covariance decreases then, as the values of one variable increase the values of the other variable fluctuate more and more, sometimes increasing, sometimes decreasing. A positive value of the covariance which is very small (i.e, which is near zero) indicates that there is no *trend* even, which would show that, though fluctuating, the values of one variable gradually increase when those of the other variable increase.

When we look at examples 12.6,12.7 and 12.10 the conclusion are similar except that here the value of the variables move in opposite directions. When the values of one variable increase those of the other decrease. If the negative value of the covariance is large in *magnitude*, as in example 12.6, it indicates that except may be for a few pairs of values an increase in the value of one variable is accompanied by a decrease in the value of the other variable and vice-versa. When the negative values of the covariance decrease in magnitude there are more and more fluctuations as described earlier till we reach a negative value near zero for the covariance, which indicates the absence of any trend.

What we have discussed above may become clearer if we make use of what is called a *scatter diagram*. If we have *n* pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ on two quantitative variables X and Y, we get the scatter diagram of the observations by plotting each pair of observations as a point in the plane as we do in coordinate geometry. The scatter diagrams of the examples given earlier are shown below.

An examination of these six scatter diagrams makes the earlier discussion easier to understand.

Scatter diagrams Fig.12.2 (a) and (b) show a clear *linear* change in the values of Y as the values of X change. Here, all the points lie on a straight line so that value of Y can be

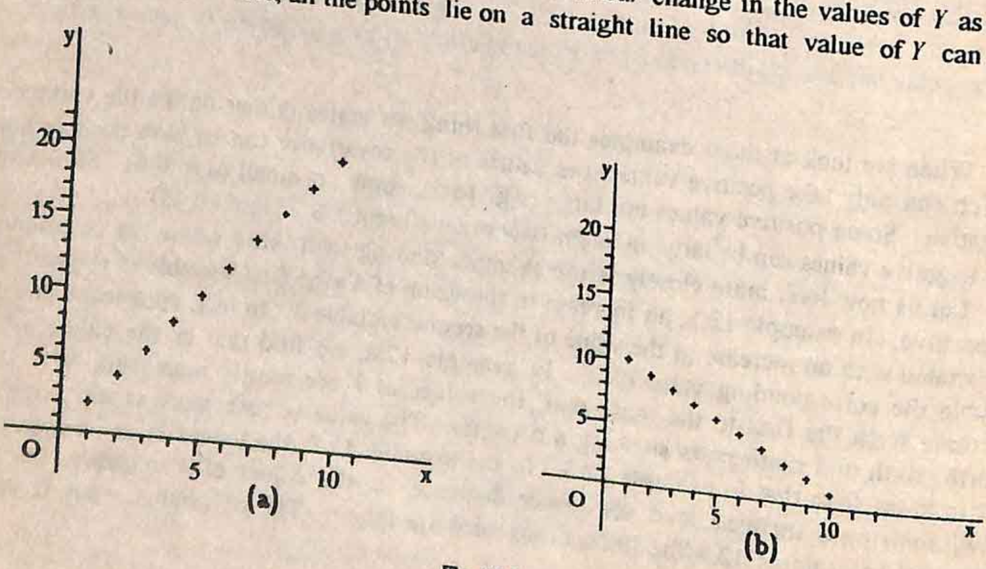


Fig. 12.2

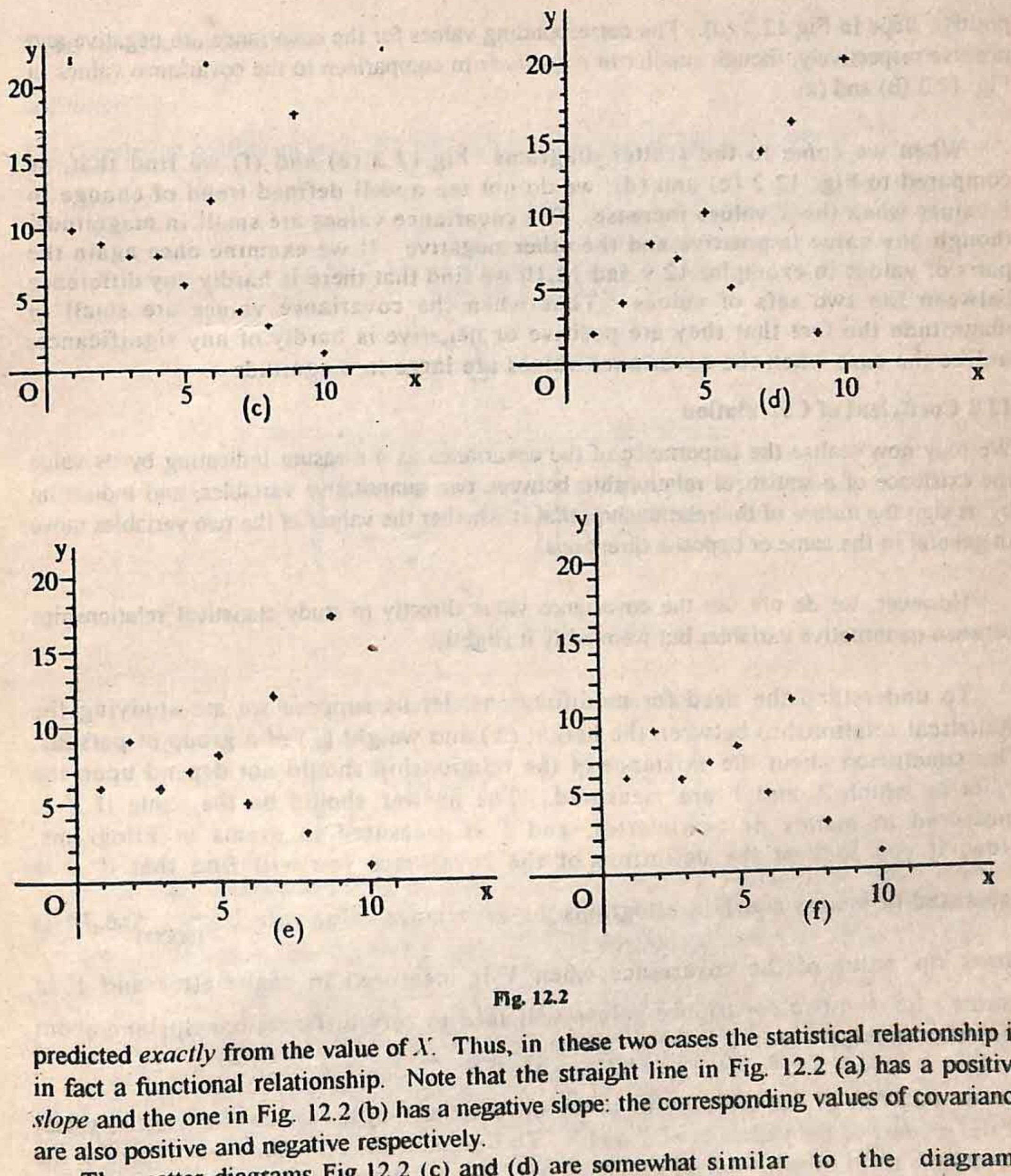


Fig. 12.2

predicted *exactly* from the value of X. Thus, in these two cases the statistical relationship is in fact a functional relationship. Note that the straight line in Fig. 12.2 (a) has a positive *slope* and the one in Fig. 12.2 (b) has a negative slope: the corresponding values of covariance are also positive and negative respectively.

The scatter diagrams Fig. 12.2 (c) and (d) are somewhat similar to the diagrams Fig. 12.2 (b) and (a) respectively except for the fact that we can no longer say that an increase in the values of one variable (say X) will *always* be accompanied by an increase (or decrease) in the values of the other variable Y. But, except for some values, we can see that the pairs of values indicate a *linear trend* with a negative slope in Fig. 12.2 (c) and a positive slope in Fig 12.2 (d). The corresponding values for the covariance are negative and positive respectively, though smaller *in magnitude* in comparison to the covariance values in Fig. 12.2 (b) and (a).

When we come to the scatter diagrams Fig. 12.2 (e) and (f) we find that, as compared to Fig. 12.2 (c) and (d), we do not see a well defined trend of change in Y values when the X values increase. The covariance values are small in magnitude, though one value is positive and the other negative. If we examine once again the pairs of values in examples 12.9 and 12.10 we find that there is hardly any difference between the two sets of values. Thus when the covariance values are small in magnitude the fact that they are positive or negative is hardly of any significance, unlike the case when the covariance values are large in magnitude.

12.8 Coeffcient of Correlation

We may now realise the importance of the covariance as a measure indicating by its value the existence of a statistical relationship between two quantitative variables, and indicating by its sign the *nature* of the relationship, that is whether the values of the two variables move in general in the same or opposite directions.

However, we do not use the covariance value directly to study statistical relationships between quantitative variables but we modify it slightly.

To understand the need for modifications, let us suppose we are studying the statistical relationship between the height (X) and weight (Y) of a group of persons. Our conclusion about the existence of the relationship should not depend upon the units in which X and Y are measured. The answer should be the same if X is measured in metres or centimetres, and Y is measured in grams or kilograms. Now, if you look at the definition of the covariance you will find that if X is measured in metres and Y in kilograms the covariance value will be $\frac{1}{100000}$(i.e., 10^{-5}) times the value of the covariance when X is measured in centimetres and Y in grams. So, the two covariance values will lead to very different conclusions about the relationship.

To take care of this difficulty we *divide* the covariance by the square root (positive value) of the *product* of the variances of X and Y. This gives us the *correlation coefficient* of the variables X and Y which we shall denote by $\rho(X, Y)$. (The letter ρ is the Greek letter "rho", which you may pronounce as "row".)

Thus, we have the definition:

Definition

The correlation coefficient $\rho(X, Y)$, between two variables X and Y is given by

$$\rho(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X) \cdot \operatorname{Var}(Y)}}\,.$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right\}}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}\left\{\sum_{i=1}^{n}(y_i - \bar{y})^2\right\}}}$$

With this modification, that is, replacing Cov (X, Y) by $\rho(X, Y)$, we get a measure which is *independent* of the units in which the quantitative variables X and Y are measured.

More precisely, if the variables X and Y are replaced by the variables $U = aX + b$, $V = cY + d$, then we have

$$\rho(U, V) = \rho(X, Y).$$

In other words, if we obtain the value of the correlation coefficient by using the values (x_i, y_i) of the variables X and Y, or by using the values $u_i = ax_i + b$, $v_i = cy_i + d$ of the variables U and V, the result is the same: we get the same value for the correlation coefficient in either case.

Note that the denominator in the definition of ρ will always be positive, so that the value of ρ is positive or negative according as the covariance value is positive or negative. Also, its value lies between -1 and $+1$, that is we have the relation (we omit the proof)

$$-1 \leq \rho(x, y) \leq 1.$$

The coefficient of correlation as defined above was first introduced by Karl Pearson and is therefore also called, "Karl Pearson's Coefficient of Correlation". It is also called the "Product Moment Correlation Coefficient" to draw attention to the use of covariance in its definition. You may also come across the name "Karl Pearson's Product Moment Correlation Coefficient"

12.9 Calculation of the Correlation Coefficient

The correlation coefficient is given by

$$\rho(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X) \cdot \operatorname{Var}(Y)}}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{\sum (x_i - \bar{x})^2\right\}\left\{\sum (y_i - \bar{y})^2\right\}}}$$

However, we do not use the above formula for calculating the value of ρ. The calculations are rendered much easier by using a diffferent formula for calculation. We know already that $\Sigma (x_i - \bar{x})^2$ and $\Sigma (y_i - \bar{y})^2$ are more easily calculated as

$$\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \text{ and } \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}.$$

In the same way the term $\Sigma (x_i - \bar{x})(y_i - \bar{y})$ is more easily calculated as

$$\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}.$$

Thus, to calculate ρ we should use the formula

$$\rho(X, Y) = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sqrt{\left\{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right\}\left\{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right\}}}$$

In this way computational labour will be less and errors of computation and rounding off will be reduced.

Example 12.11

Let us take the data of Example 12.3 :

(1, 10), (2, 9), (3, 8), (4, 8), (5, 6),
(6, 12), (7, 4), (8, 3), (9, 18), (10, 1)

We have

$$\sum x_i = 55, \quad \sum y_i = 79$$

$$\sum x_i^2 = 1 + 4 + 9 + \ldots + 81 + 100 = 385$$

$$\sum y_i^2 = 100 + 81 + 64 + \ldots + 1 = 839$$

$$\sum x_i y_i = 10 + 18 + \ldots + 162 + 10 = 410$$

$$n = 10$$

Hence,

$$\rho(X, Y) = \frac{410 - \dfrac{55 \times 79}{10}}{\sqrt{\left(385 - \dfrac{55 \times 55}{10}\right)\left(839 - \dfrac{79 \times 79}{10}\right)}}$$

$$= \frac{-24.5}{\sqrt{(82.5)(214.9)}} = \frac{-24.5}{133.15}$$

$$= -0.18 \text{ (approx.)}.$$

Example 12.12

We consider the same example but change the values x_i to $u_i = 2x_i + 1$ and y_i to $v_i = y_i - 5$. The data now is

(3, 5), (5, 4), (7, 3), (9, 3), (11, 1),
(13, 7), (15, −1), (17, −2), (19, 13), (21, −4).

For this data we have

$$\sum u_i = 120, \quad \sum v_i = 29$$

Hence, $\sum u_i^2 = 1770, \quad \sum v_i^2 = 299, \quad \sum u_i v_i = 299.$

$$\rho(U, V) = \frac{299 - \dfrac{120 \times 29}{10}}{\sqrt{\left(1770 - \dfrac{120 \times 120}{10}\right)\left(299 - \dfrac{29 \times 29}{10}\right)}}$$

$$= \frac{-49}{\sqrt{(330)(214.9)}} = \frac{-49}{266.302}$$

$$= -0.18 \text{ (approx.)}$$

12.10 Interpretation of the Correlation Coefficient

The coefficient of correlation is designed to give us a measure indicating by its value if a statistical relationship exists or not between two quantitative variables. We shall now discuss how the value of the coefficient of correlation enables us to decide if a statistical relationship exists or not. We begin by obtaining the value of the correlation coefficient for the data of Examples 12.5 to 12.10. We give below the values of Cov (X, Y), which we have already obtained, Var (X), Var (Y), and $\rho(X, Y)$ for each of the six Examples 12.13 to 12.18.

Example 12.13

$$\text{Cov}(X, Y) = 16.5, \quad \text{Var}(X) = 8.25, \quad \text{Var}(Y) = 33$$
$$\rho(X, Y) = 1$$

Example 12.14

$$\text{Cov}(X, Y) = -8.25, \quad \text{Var}(X) = 8.25, \quad \text{Var}(Y) = 8.25$$
$$\rho(X, Y) = -1$$

Example 12.15

$$\text{Cov}(X, Y) = -2.45, \quad \text{Var}(X) = 8.25, \quad \text{Var}(Y) = 21.49$$
$$\rho(X, Y) = -0.18$$

Example 12.16

$$\text{Cov}(X, Y) = 10.2, \quad \text{Var}(X) = 8.25, \quad \text{Var}(Y) = 33.96$$
$$\rho(X, Y) = 0.61$$

Example 12.17

$$\text{Cov}(X, Y) = 0.4, \quad \text{Var}(X) = 8.25, \quad \text{Var}(Y) = 18.64$$
$$\rho(X, Y) = 0.03$$

Example 12.18

$$\mathrm{Cov}\,(X, Y) = -\,0.45, \quad \mathrm{Var}\,(X) = 8.25, \quad \mathrm{Var}\,(Y) = 14.29$$
$$\rho\,(X, Y) = -\,0.04$$

(Values of ρ have been rounded off to two decimal places in Examples 12.15 to 12.18)

Examples 12.13 and 12.14 result in the extreme values 1 and -1 of the correlation coefficient. We also note that in these two examples the scatter diagram is a straight line showing that there is a functional relationship, in fact a linear relationship, between the two variables. If the value of one variable is known for a given unit of observation the value of the other variable for the same unit of observation is exactly determined. The extreme values 1 and -1 of the correlation coefficient arise only in such situations. We state without proof the following result:

"The correlation coefficient $\rho\,(X, Y)$ takes the value 1 or -1 if and only if the variables X and Y are linearly related".

Consider next the positive values of $\rho\,(X, Y)$. As the value of ρ decreases from 1 to 0 we find that the scatter digram indicates a linear *trend* with the points being scattered around a straight line. The scatter is small for positive values of ρ close to 1 and increases as the value of ρ decreases; for values of ρ near 0 the scatter is considerable and the linear trend almost disappears. The situation is similar in the case of negative values of the correlation coefficient. As the values of ρ increase from -1 to 0 the scatter diagram continues to show a linear trend with small scatter for values of ρ near -1 and increasing scatter as the values of ρ approach 0. For negative values of ρ near 0 the scatter is again considerable and the linear trend disappears. The positive and negative signs of the value of the correlation coefficient only indicate whether the values of the two variables increase or decrease together (positive values) or, the values of one variable decrease with increase in the values of the other variable (negative values).

We thus see that the value of the correlation coefficient enables us to find out if a statistical relationship exists between variables in a bivariate frequency distribution. Once we know the value of the correlation coefficient we can draw some conclusions regarding the existence of a statistical relationship between two variables on the basis of the observations we have on the two variables. We cannot go into the complete mathematical theory at this stage, but the general rules and methods of interpretation of the value of the correlation coefficient can be stated as follows:

(1) If the value of the correlation coefficient is equal to $+1$, then there is necessarily a *functional* relationship between the variables X and Y. In fact, we can say that the variables have a *linear* relationship, i.e., there exist constants a and b (≥ 0) such that

$$Y = a + bX.$$

(It automatically implies that there exist constants c and d (≥ 0) such that

$$X = c + dY\ )$$

The scatter diagram in this case is a straight line with a positive slope. In such a case we sometimes say that there is perfect positive correlation between X and Y.

(2) If the value of the correlation coefficent is equal to -1, then there is necessarily a functional linear relationship between X and Y. There exist coefficients a, b or c, d such that

$$Y = a + bX$$

or,
$$X = c + dY$$

However, in this case b and d will be negative, whereas in the first case ($\rho = 1$) they were positive. The scatter diagram in this case is a straight line with a negative slope. In this case we say that there is perfect negative correlation between X and Y.

(3) If the value of $\rho(X, Y)$ is positive and close to the value 1, the points in the scatter diagram are scattered about a straight line with a positive slope. The points will be close to the straight line if the value of ρ is very near to 1. We say in such case that there is high positive correlation between X and Y.

(4) If $\rho(X, Y)$ has a negative value which is close to -1 the situation is similar to the preceding case. The points of the scatter diagram now are scattered around a straight line with a negative slope; they get closer and closer to the straight line as the (negative) value of ρ gets near to -1. We say in such cases that there is high negative correlation between X and Y.

(5) When the value of ρ (positive or negative) approaches the value 0 we are faced with a difficulty. In such cases either there is no evidence of a statistical relationship between the variables, or the variables can even be having a functional relationship. Of course, the functional relationship in such cases will not be linear since for linear functional relationship we must have $\rho = 1$ or $\rho = -1$.

So, when we have a value of ρ near 0 we cannot draw any conclusion about the existence of a statistical relationship between the variables unless we draw the scatter diagram. When we have a value of ρ near to 1 or -1, we do not have to draw the scatter diagram, the value of ρ itself tells us what the scatter diagram will look like. But when the value of ρ is near 0 we must look at the scatter diagram before we draw any conclusions. The scatter diagram alone will tell us whether the small value of ρ represents the absence of a statistical relationship (in which case we will say that the variables are *uncorrelated*), or it represents the existence of a *non-linear* functional or statistical relationship.

The correlation coefficent is derived from the numerical values of the given variables but does not in any way depend on what the variables are. Hence, care must be taken in interpreting an observed value of the correlation coefficent. A high value of the correlation coefficient should not automatically lead one to the conclusion that the variables have some influence on each other. For example, a value of -0.98 has been reported for the correlation between the birth rate in Great Britain, from 1875 to 1920, and the production of pig iron in the United States. Surely, we cannot, by any stretch of imagination, conclude from it that the production of pig iron and the birth rate influence each other in any way. Similarly, a high value of correlation between rainfall and wheat production may be taken to support the conclusion that rainfall influences wheat production, but it certainly cannot be interpreted to mean that wheat production influences rainfall. Thus, the correlation coefficient by itself does not say anything about the influence that one variable has on another. At best, it can only support or disprove our hypothesis of two variables influencing each other, but the hypothesis has to be formulated by taking the nature of the variables into consideration.

C. REGRESSION ANALYSIS

12.11 Introduction

At the start of this chapter we introduced the idea of statistical reationship between variables as distinct from the idea of a functional relationship. In a functional relationship, if we know the value of one variable (the independent variable) the value of the other variable (the dependent variable) is determined exactly. In a statistical relationship the situation is slightly different. Now we cannot exactly determine the value of one variable from that of the other. All we can do is to make an estimate of the value of one variable when that of the other is known, knowing fully well that there could be an error as there is no certainty that our estimate would tally exactly with the value actually observed. *Regression analysis* is concerned with the method of making such estimates. It attempts to lay down rules for *predicting* the value of one variable from that of the other such that our prediction will be "good" in some sense.

Actual prediction is done through a rule which lays down what the *predicted value* of a variable will be for a *given value* of the other variable. In other words, we will define a function f of the X- values so that if x is the given value of the variable X the predicted value of the variable Y is $f(x)$. Similarly, we define another function g of the Y values which gives the predicted value $g(y)$ of the variable X for a given value y of the variable Y.

We use the general term *regression function* for the functions f and g and say that

$$y = f(x)$$

is the *regression equation of Y on X*, and

$$x = g\ (y)$$

is the *regression equation of X on Y*. Thus, a regression equation describes the rule to be followed for determining the predicted value of one variable from the given value of the other variable. There are *two* regression equations. The regression equation of Y on X is used to predict the values of Y from the given values of X, and the regression equation of X on Y is used to obtain the predicted value of X from the given value of Y.

Once the regression rule has been defined we have to somehow try to see how good it is. Suppose our regression rule is defined by the regression equation

$$y = f(x)$$

of Y on X. If (x_i, y_i) are the observed values of X and Y for the ith unit of observation, then the difference between the actual value y_i of Y and the predicted value $f\ (x_i)$ gives an idea of the *error* made in predicting the value of Y from the given value x_i of the variable X. If the differences

$$y_1 - f(x_1), y_2 - f(x_2), \ldots, y_n - f(x_n)$$

are small, we could say that our prediction rule is "good". Thus, a measure of the error we will be making by using the prediction rule described by the regression equation will have to take into account all the above differences. You may have noticed that the situation is similar to that of constructing a measure of dispersion of the values out of the individual deviations

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}.$$

which was discussed in Chapter 17 of the textbook for Class XI.

The general regression problem has two aspects, one is the selection of the regression function, and the other is the construction of a measure of the prediction error made while using the regression function for prediction purposes. We shall discuss both these aspects, limiting ourselves to the case where the regression function is linear. In other words, we will be studying the method of *linear regression* only.

There is another reason for restricting ourselves to the study of linear regression. One can try to tackle the problem of prediction only when one is reasonably sure that a statistical relationship exists between the variables. In the absence of such a relationship one cannot think of predicting the values of one variable from those of the other. Such a prediction is meaningless. It would be like predicting, for example, the amount of wheat produced in a year from the number of train accidents during that year. We have used the correlation coefficient to determine if a statistical relationship exists between the variables. And we have seen that the correlation coefficient can be used only when the relationship can be assumed to be of a linear type. Thus, our prediction methods will also have to use a linear form for the regression function. If any other form is to be used then the methods for determining the existence of a statistical relation will also have to be changed.

12.12 Least Squares:Lines of Regression

Suppose we wish to predict the values of the variable Y from the values of the variable X. We take as our regression equation, the equation

$$y = a + bx.$$

Thus, if x_i is an observed value of X, the predicted value of Y for the given value of X will be

$$a + bx_i.$$

We are thus restricting ourselves to the use of linear regression functions. The problem now is to determine the constants a and b in such a manner that the particular linear function determined by these values of a and b is, in some sense, the best among all linear functions. The actual values of a and b will depend on the observations $(x_1, y_1), \ldots, (x_n, y_n)$ on the variables X and Y. As mentioned earlier the best linear function will be obtained by examining the differences

$$y_1 - (a + bx_1), y_2 - (a + bx_2), \ldots, y_n - (a + bx_n)$$

between the observed values

$$y_1, y_2, \ldots, y_n$$

and the predicted values

$$a + bx_1, a + bx_2, \ldots, a + bx_n$$

of the variable Y. The best linear function will be the one for which the total contribution of these differences is small. Now, the predicted value will be sometimes more and sometimes less than the observed value so that, some of the above differences will be positive and some negative. Hence, we cannot simply total them and then choose a and b such as to make this total as small as possible. For this reason, we take instead the sum

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2$$

of the squares of these differences and then choose a and b such that this sum is a minimum. The values of a and b can be easily determined by the methods of the differential calculus for obtaining the maximum and minimum of functions. Once the values of a and b have been determined, the regression equation takes the simple form

$$y - \bar{y} = \frac{\operatorname{Cov}(X, Y)}{\sigma_x^2} (x - \bar{x}),$$

where σ_x^2 denotes the variance of X. This equation is the equation of the *least squares line of regression of Y on X*. The constant

$$\frac{\operatorname{Cov}(X, Y)}{\sigma_x^2} = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

is called the *regression coefficient of Y on X* and is denoted by b_{yx}. Since the value of the variance σ_x^2 is positive, the regression coefficient will be positive when Cov (X, Y) is positive, and it will be negative when Cov (X, Y) is negative. Thus, the regression coefficient b_{yx} has the same sign as the correlation coefficient $\rho(X, Y)$.

Similarly, if we wish to obtain the *least squares line of regression of X on Y*, we will start with the regression equation

$$x = c + d\,y$$

and minimise the sum

$$\sum_{i=1}^{n} (x_i - c - dy_i)^2.$$

The values of c and d which give the minimum value of this sum result in the following form for the regression line X on Y:

$$x - \bar{x} = \frac{\mathrm{Cov}\,(X, Y)}{\sigma_y^2}\,(y - \bar{y}).$$

where σ_y^2 denotes the variance of Y. The constant

$$\frac{\mathrm{Cov}\,(X, Y)}{\sigma_y^2} = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum (y_i - \bar{y})^2}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2}$$

is called the *regression coefficient of X on Y* and is denoted by b_{xy}. It also has the same sign as the covariance of X and Y, or as the correlation coefficient $\rho\,(X, Y)$.

If we wish to predict the value of Y for a given value x_i of X, we will use the regression line of Y on X to get the predicted value of Y as

$$\bar{y} + b_{yx}\,(x_i - \bar{x})$$

Similarly, for a given value y_i of Y, the predicted value of X will be obtained by using the regression line of X on Y, and will be given by

$$\bar{x} + b_{xy}\,(y_i - \bar{y}).$$

Note that the regression lines can be used for prediction even when the given values x_i, or y_i, are not actually observed but are given otherwise.

The lines of regression obtained above are called the least squares lines of regression because they have been obtained by minimising the sums of squares

$$\sum (y_i - a - bx_i)^2, \quad \sum (x_i - c - dy_i)^2.$$

The words "least squares" are used to indicate this. The words, "Y on X" and "X on Y" are used to indicate that the regression line is to be used for predicting respectively the values of Y and of X.

12.13 Calculation of Regression Coefficients

We have already seen that the covariance of X and Y is most easily and simply calculated as

$$\frac{1}{n}\left[\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}\right]$$

and that for the variance we should use the expression

$$\frac{1}{n}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]$$

Hence, the calculation of the regression coefficients b_{yx} and b_{xy} should be made by using the expressions

$$b_{yx} = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}$$

and

$$b_{xy} = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}$$

If these expressions are used, not only will the labour of calculation be reduced but also the computational and rounding off errors.

Example 12.19

Consider the observations

(1, 2), (2, 4), (3, 8), (4, 7), (5, 10),
(6, 5), (7, 14), (8, 16), (9, 2), (10, 20)

for which the correlation coefficient was found to be high (=0.61). We have

$$\sum x = 55, \ \sum y = 88, \ \bar{x} = 5.5, \ \bar{y} = 8.8,$$

$$\sum x^2 = 385, \ \sum y^2 = 1114, \ \sum xy = 586.$$

Hence,

$$b_{yx} = \frac{586 - \dfrac{55 \times 88}{10}}{385 - \dfrac{55 \times 55}{10}} = \frac{102}{82.5} = 1.24 \text{ (approx.)}$$

and

$$b_{xy} = \frac{586 - \dfrac{55 \times 88}{10}}{1114 - \dfrac{88 \times 88}{10}} = \frac{102}{339.6} = 0.30 \text{ (approx.)}$$

The regression line of Y on X is

$$(y - 8.8) = (1.24)(x - 5.5)$$

or, $$y = 1.24x + 1.98$$

The regression line of X on Y is

$$(x - 5.5) = (0.3)(y - 8.8)$$

or, $$x = 0.3y + 2.86$$

Using these regression lines the predicted value of Y, given that the X – value is 6.5, is

$$y = (1.24)(6.5) + 1.98$$
$$= 10.04$$

and the predicted value of X, given that the Y-value is 9, is

$$x = (0.3)(9) + (2.86)$$
$$= 5.56$$

EXERCISE 12.2

1. Draw the scatter diagram of the bivariate data of Example 12.19 and draw the lines of regression on it.

12.14 Relationship with Correlation Analysis

As mentioned earlier, regression analysis, that is, the prediction procedure, makes sense only when correlation analysis has established the existence of a statistical relationship between the variables. However, the relationship between correlation and regression analysis goes much deeper as we will now try to show.

We begin with the study of the amount of error resulting from the use of the least squares lines of regression to predict the value of one variable from that of the other. Suppose we are predicting the value of Y. Then the error of prediction is measured by the sum

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2$$

of the squares of the differences between the observed values y_i and predicted values $a + bx_i$ of the variable Y. Using the values of a and b the measure of error is given by

$$\sum_{i=1}^{n} \{ y_i - \bar{y} - b_{yx} (x_i - \bar{x}) \}^2$$

$$= \sum (y_i - \bar{y})^2 + b_{yx}^2 \sum (x_i - \bar{x})^2 - 2b_{yx} \sum (y_i - \bar{y})(x_i - \bar{x})$$

$$= n\sigma_y^2 + \frac{\{ \text{Cov}(X, Y) \}^2}{\sigma_x^4} n\sigma_x^2 - 2 \frac{\text{Cov}(X, Y)}{\sigma_x^2} n \, \text{Cov}(X, Y)$$

$$= n \left[\sigma_y^2 - \frac{\{ \text{Cov}(X, Y) \}^2}{\sigma_x^2} \right]$$

$$= n\sigma_y^2 \left[1 - \frac{\{ \text{Cov}(X, Y) \}^2}{\sigma_x^2 \sigma_y^2} \right]$$

$$= n\sigma_y^2 (1 - \rho^2).$$

Thus we see immediately that the error of prediction, when the least squares line of regression is used, decreases as the value of the correlation coefficient ρ approaches 1 or -1. It becomes zero when ρ is equal to 1 or -1, meaning thereby that the observed and predicted values of Y coincide. As the value of ρ moves away from 1 or -1 towards 0 the error of prediction increases. The same is true when we are using the least squares line of regression to predict the value of X. In this case the error of prediction is given by

$$n\sigma_x^2\,(1-\rho^2)$$

We thus see that the value of the correlation coefficient ρ tells us something about the degree of error made when the least squares line of regression is used for prediction purposes.

MISCELLANEOUS EXERCISE ON CHAPTER 12

1. The marks obtained by 100 students in tests in Mathematics and Physics were as follows. The first number in each order pair denotes marks in Mathematics

(13,12), (5,11), (13,21), (21,29), (62,57), (54,58), (53,28), (58,19), (43,68), (31,28),
(26,11), (18,32), (29,34), (37,21), (43,40), (54,17), (56,59), (61,29), (19,20), (8,7),
(6,27), (28,22), (38,29), (70,68), (69,64), (62,34), (41,33), (27,19), (24,35), (8,13),
(6,3), (32,20), (29,17), (48,62), (63,23), (55,34), (23,39), (11,23), (11,19), (6,8),
(37,29), (55,43), (57,36), (43,29), (26,13), (35,38), (17,14), (58,23), (54,36), (64,40),
(7,4), (20,18), (0,3), (28,19), (68,39), (34,37), (4,9), (31,20), (56,67), (48,39),
(30,18), (52,18), (38,55), (14,18), (63,61), (30,27), (61,37),(54,41), (30,37), (17,27),
(32,39), (39,18), (43,27), (67,62), (31,38), (40,19), (28,9), (26,14), (97,95), (68,79),
(88,68), (83,67), (72,38), (91,63), (93,69), (78,94), (92,57), (81,58), (73,40), (65,83),
(79,84), (63,45), (93,48), (77,63), (76,62), (68,59), (74,59),(90,96), (89,93), (21,18).

(a) Present the data in the form of a bivariate frequency table dividing the Mathematics marks into 5 classes 0-25, 26-50, 51-70, 71-90 and 91-100, and the Physics marks into 4 classes 0-20, 21-40, 41-70 and 71-100. (Use the method of tally marks.)

(b) Obtain the marginal distributions for marks in Mathematics and in Physics.

(c) Write down the conditional distribution of the students by marks in Mathematics given that the marks in Physics were at least 41.

(d) Find the conditional distribution of the students by marks in Physics given that the marks in Mathematics were between 21 and 70.

(e) How many students were there whose marks in Mathematics were at most 70 and marks in Physics were at least 41?

(f) Find the number of students whose marks in Mathematics were between 26 and 90 and marks in Physics were less than or equal to 40.

2.(a) Calculate the value of the correlation cofficient for the following data:

(1,13), (2,23), (3,33), (4,43), (5,53),
(6.63), (7,73), (8,83), (9,93), (10,103),
(11.10.5), (12,11), (13,11.5), (14,12), (15, 12.5),
(16.13), (17,13.5) (18, 14), (19,14.5), (20,15).

(b) Draw the scatter diagram

(c) Comment on the results.

3. The data below gives the values of (X, Y) for 16 persons where X is the age in years and Y the weight in kilograms:

(25,55), (30,62), (30,61), (30,63),
(35,67), (40,70), (40,71), (45,71),
(45,72), (50,70), (55,67), (55,65),
(55,68), (60,62), (60,61), (65,55)

(a) Calculate the value of the correlation coefficient.

(b) Draw the scatter diagram.

(c) Comment on the results.

4. The data below gives the value of the annual sales (X) and the annual expenditure (Y) on advertising of a firm for the years 1972 to 1982

Year	X (Rs in 000's)	Y (Rs in 000's)
1972		20
1973	34	21
1974	40	17
1975	27	9
1976	17	22
1977	36	7
1978	19	23
1979	38	16
1980	23	5
1981	11	14
1982	21	–

(a) Calculate the value of the correlation coefficient between annual sales and advertisement expenditure.

(b) Predict the sales for the year 1978 from the advertisement expenditure of 1978 and compare with the actual sales.

(c) Calculate the coefficient of correlation between yearly advertisement expenditure and the annual sales in the succeeding year.

(d) Predict the sales for the year 1977 from the advertisement expenditure of 1976 and compare with the actual sales.

5. (a) If the values x_i and y_i are transformed to u_i and y_i by means of the relations

$$u_i = ax_i + b$$
$$v_i = cy_i + d$$

prove that

$$\rho(X, Y) = \rho(U, V).$$

(b) Find the relation between b_{yx} and b_{vu}, and between b_{xy} and b_{uv}.

6. Prove that

$$b_{yx} \cdot b_{xy} = [\rho(X, Y)]^2$$

7. (a) Find the angle between the regression lines of Y on X and X on Y.

[**Hint**: The slopes are b_{yx} and $\dfrac{1}{b_{xy}}$.]

(b) From the above prove that there is only one regression line (that is, the two lines of regression coincide) if $\rho(X, Y) = \pm 1$.

8. Find Cov (X, Y) between X, Y, if

X	1	2	3	4	5
Y	2	4	6	8	10

9. Find Karl Pearson's coefficient of correlation between X and Y for the following data:

X	2	3	4	5	6
Y	4	3	2	8	10

10. Find the regression coefficient b_{yx} between X and Y for the following data:

$\sum x = 24$, $\sum y = 44$, $\sum xy = 306$, $\sum x^2 = 164$,

$\sum y^2 = 574$, $N = 4$

11. Out of the following two regression lines find the line of regression of Y on X:

$$x + 2y - 5 = 0$$
$$2x + 3y = 8$$

12. For the following observations, find the regression line of Y on X in the form

$$y - \bar{y} = b_{yx}\,(x - \bar{x})$$

$$\{(x, y)\} = \{(1, 7), (4, 5), (7, 2), (10, 6), (13, 3)\}$$

13. $4x + y - 10 = 0$, $2x + 5y - 14 = 0$ are two regression lines. Find the correlation coefficient between X and Y.

14. Estimate the two lines of regression by the least squares method for the following:

X	1	2	3
Y	2	4	5

CHAPTER 13

Computing

13.1 Introduction

Current period is described by calling it the age of computers. As days pass by, computers are being used in different walks of life. Computer revolution is expected to make more lasting, all pervasive and widespread impact on human life and civilization. In your earlier classes you have studied some aspects related to computing*. Names of many mathematicians are associated with the history of development of computers and with the theories of computing. Pascal, Leibnitz (who invented calculus independently of Newton), Charles Babbage, Alan Turing, John von Neumann are some of them.

13.2 Block Diagram of a Computer

A computer has different components. We represent these by blocks in a diagram and give a block diagram of a computer in Fig.13.1.

Major components of a computer, as shown, are (1) input unit, (2) memory (or storage, or random access memory), (3) output unit, (4) control unit, and (5) arithmetic logical unit (ALU). Some computers, in addition, have auxiliary memory in the form of magnetic tapes, discs, etc. A computer is also called an information processing unit.

Information consisting of data and instructions is taken in by the input unit from the users and is passsed on to memory for storage and onward use. ALU is the unit where all arithmetic and logical (e.g. comparing two numbers to decide which one is bigger) computations are carried out. For carrying out these computations there are registers in the ALU. Users get all the computed results from the output unit. Control unit, as the name suggests, is the one which controls the whole working of the computer by sending electronic command signals to other components of the computer. Control unit and ALU taken together is called *central processing unit* (CPU).

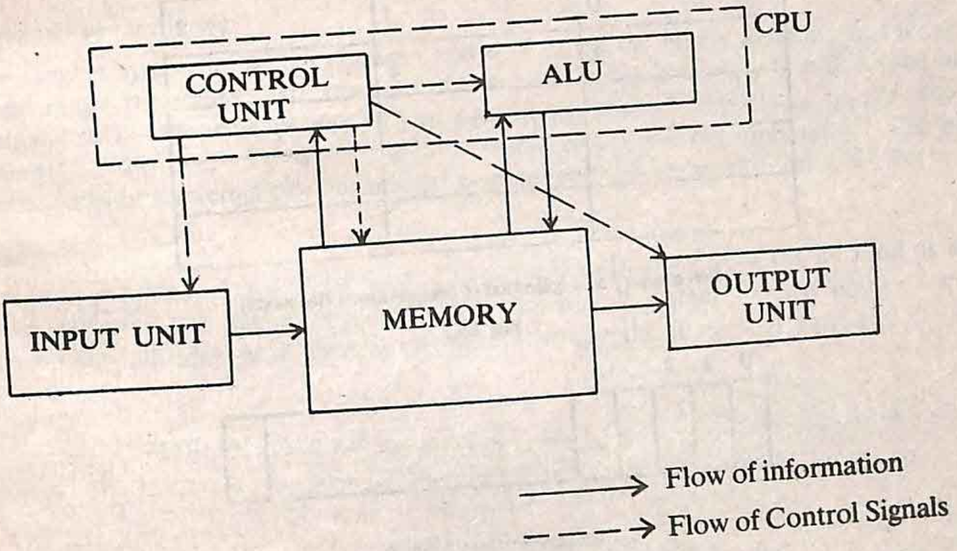
A computer may have more than one input and output units. For example, printer and display screen are two different output units attached to the same computer.

Remark

A present day computer is an electronic, digital machine in which *stored-program concept* is incorporated. An algorithm (in the form of a program written using a programming language)

---

*Students will be benefited by referring to Chapter 17, in Class IX book and to Chapter 12, in Class X book, on computing.

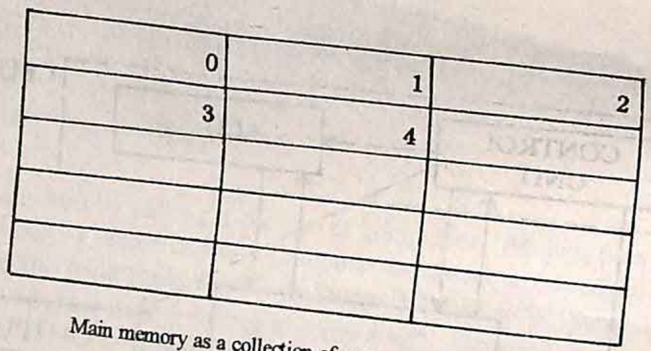


Block diagram of a computer
Fig. 13.1

and the data with which it is to be executed are both stored in the memory of a computer. This was a revolutionary idea. Credit of this idea goes to John von Neumann and Alan Turing. Storing of algorithm in memory is necessary if a series of instructions is to be repetitively carried out. Early machine ENIAC (Electronic Numerical Integrater and Computer) did not employ the stored-program concept in it. In ENIAC only data was stored in a memory and the algorithm used to be in other outside panel. Stored program concept is an integral characteristic of present day computers. In today's terminology, ENIAC is not a computer, but a calculator.

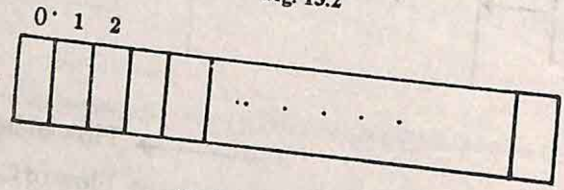
13.3 Memory

Our aim is to see how we can use the computer to solve some problems. For that purpose, it is useful to know a little more about main memory. From the users' point of view, main memory can be thought of as a collection of compartments (or locations), as shown in Fig. 13.2. Each compartment is assigned a number called its address (starting with zero as shown in the Fig. 13.2). The total number of compartments gives us the size of the memory.



Main memory as a collection of compartments (locations)

Fig. 13.2



Bits in a memory location

Fig. 13.3

Each compartment of memory (as well as a register in ALU) consists of sub-compartments (See Fig. 13.3). Each sub-compartment can store either a zero or a 1. Any information to be stored inside a computer is put using zeros and 1's. The digits 0 and 1 are called binary digits (bits in short). The acronym **bit** is formed by taking the letter *b* from the word 'binary' and the letters *i*, *t* from the word 'digit'. Similarly, we have the acronym **dit** for decimal digit, **hit** for hexadecimal digit, etc. The number system that uses only two digits is called binary number system. Computers use binary number system for computation. The computations in binary and other related number systems are discussed in the next section.

13.4 Number Systems .

In the number system which we use in our daily life (known as decimal system), we use ten basic symbols called digits namely 0, 1, 2,..., 9 and with the help of these 10 digits we are able to write any rational number. The decimal system is a *place-value* system, meaning thereby that the value represented by a digit depends upon the place of the digit within the numeral. The values assigned to consecutive places in the decimal system are $\ldots, 10^4, 10^3, 10^2, 10^1, 10^0, 10^{-1}, 10^{-2}, 10^{-3}, \ldots,$ (from left to right). Thus

$$2742.35 = 2 \cdot 10^3 + 7 \cdot 10^2 + 4 \cdot 10^1 + 2 \cdot 10^0 + 3 \cdot 10^{-1} + 5 \cdot 10^{-2}$$

As ten basic symbols are used for representing the numbers, ten is called base (or radix) of the system and the system is called 'base-ten' system or decimal system.

Binary Number System

The number system for which the base is two is called the binary system. In this system numbers are represented with the help of two basic symbols namely 0 and 1. The values assigned to consecutive places in the system are (when expressed in the decimal system),..., 2^4, 2^3, 2^2, 2^1, 2^0, 2^{-1}, 2^{-2},. . . where 2^0 place is the units place. The binary numeral can be converted into the decimal numeral and vice versa as illustrated below.

Example 13.1

Convert $(1101101)_2$ into decimal form. (The 2 written by the side of the numeral as shown above indicates that the numeral belongs to the base 2 system. When numeral appears without any indication of its base, by convention we take the numeral as a decimal numeral.)

Solution

$$\begin{aligned}(1101101)_2 &= 1 \cdot 2^6 + 1 \cdot 2^5 + 0 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 \\ &= 64 + 32 + 0 + 8 + 4 + 0 + 1 \\ &= 109\end{aligned}$$

Example 13.2

Convert 109 into a binary numeral.

Solution

Divide 109 by 2. You get 54 as quotient and 1 as remainder. Again divide 54 by 2. You get 27 as quotient and 0 as remainder. Continue the process till you have 0 as the quotient . See the working below. Write the remainders you get from bottom to top in a row from left to right. This is the binary representation of 109.

$$\begin{array}{r|rcl} 2 & 109 & & \\ \hline 2 & 54 & \cdots & 1 \\ \hline 2 & 27 & \cdots & 0 \\ \hline 2 & 13 & \cdots & 1 \\ \hline 2 & 6 & \cdots & 1 \\ \hline 2 & 3 & \cdots & 0 \\ \hline 2 & 1 & \cdots & 1 \\ \hline & 0 & \cdots & 1 \end{array} \Bigg\uparrow \qquad 109 = (1101101)_2$$

Example 13.3

Convert $(1011.101)_2$ into the decimal form.

Solution

$$\begin{aligned}(1011.101)_2 &= 1\cdot 2^3 + 0\cdot 2^2 + 1\cdot 2^1 + 1\cdot 2^0 + 1\cdot 2^{-1} + 0\cdot 2^{-2} + 1\cdot 2^{-3}\\ &= 8 + 0 + 2 + 1 + 1\cdot(.5) + 0\cdot(.25) + 1\cdot(.125)\\ &= (11.625)_{10}\end{aligned}$$

Note: The point appearing between 1011 and 101 in the binary numeral is called the binary point, while the point in 11.625 is called the decimal point.

Example 13.4

Convert 11.625 into the binary form.

Solution

When we have a decimal numeral having both integral and fractional parts, then we convert the integral part and the fractional parts separately. You know how to convert 11 into binary form which is 1011. While converting the integral part we divide the given number and subsequent quotients by 2. To convert the fractional part, we will multiply the fractional part by 2 and obtain the product. We take the integral part of the product out and multiply the fractional part by 2. We proceed in the same way until we get all zeros in the fractional part or till we get required number of binary digits. The procedure for converting the fractional part is shown below:

$$\left\downarrow\begin{array}{rcr} & & 0.625\\ & & \times\ 2\\ \hline 1 & \cdots & \cancel{1}.250\\ & & \times\ 2\\ \hline 0 & \cdots & \cancel{0}.500\\ & & \times\ 2\\ \hline 1 & \cdots & \cancel{1}.000\end{array}\right.$$

$$0.625 = (.101)_2$$

$$11.625 = (1011.101)_2$$

Note: It is not necessary that the binary representation of a terminating decimal is also terminating as can be seen from the following example:

0.1 is a terminating decimal.

Let us convert this into binary system.

$$\begin{array}{rr@{\qquad\qquad}rr}
 & 0.1 & 1\ \cdots & 1.6 \\
 & \underline{\times\ 2} & & \underline{\times\ 2} \\
0\ \cdots & 0.2 & 1\ \cdots & 1.2 \\
 & \underline{\times\ 2} & & \underline{\times\ 2} \\
0\ \cdots & 0.4 & 0\ \cdots & 0.4 \\
 & \underline{\times\ 2} & & \underline{\times\ 2} \\
0\ \cdots & 0.8 & 0\ \cdots & 0.8 \\
 & \underline{\times\ 2} & &
\end{array}$$

Thus, the cycle continues.

$$\therefore \quad (0.1)_{10} = (0.0001100110\cdots)_2 \text{ (non-terminating)}$$

Octal Number System

As the name implies this is base eight (2^3) system. The numerals are written with the help of eight basic symbols namely 0, 1, 2, . . . , 7. The value (expressed in the decimal system) assigned to consecutive places are. . . , 8^3, 8^2, 8^1, 8^0, 8^{-1}, 8^{-2}, . . . , where 8^0 place is the unit's place. The procedures for converting a decimal numeral into an octal numeral and the other way round are similar to the procedures discussed in connection with binary system. The following examples illustrate the procedures.

Example 13.5

Convert $(1632.23)_8$ into decimal form.

Solution

$$\begin{aligned}(1632.23)_8 &= 1 \times 8^3 + 6 \times 8^2 + 3 \times 8^1 + 2 \times 8^0 + 2 \times 8^{-1} + 3 \times 8^{-2} \\ &= 1 \times 512 + 6 \times 64 + 3 \times 8 + 2 \times 1 + 2 \times (0.125) + 3 \times (0.015625) \\ &= 512 + 384 + 24 + 2 + 0.250 + 0.046875 \\ &= 922.296875\end{aligned}$$

Example 13.6

Convert 922.296875 into octal system.

Solution

We have to convert the integral part and fractional part separately. The procedure is shown below:

$$\begin{array}{rl|l} 8 & 922 & \\ \hline 8 & 115 \ \ldots & 2 \\ \hline 8 & 14 \ \ldots & 3 \\ \hline 8 & 1 \ \ldots & 6 \\ \hline & 0 \ \ldots & 1 \end{array} \uparrow \qquad \downarrow \begin{array}{rr} & 0.296875 \\ & \times 8 \\ \hline 2 \ \ldots & \cancel{2}.375000 \\ & \times 8 \\ \hline 3 \ \ldots & \cancel{3}.000000 \end{array}$$

$$\therefore \quad 922.296875 = (1632.23)_8$$

Hexadecimal System

As the name implies, this is the base sixteen system. The numerals are written with the help of sixteen symbols, namely 0, 1, 2, 3 ,4, 5, 6,7, 8, 9, *A*, *B*, *C*, *D*, *E*, *F*. Note that the symbol *A* represents the number ten (in the decimal system). Similarly *B*, *C*, *D*, *E*, *F* represent respectively the numbers eleven, twelve, thirteen, fourteen and fifteen. The values assigned to consecutive places are . . . , 16^2, 16^1, 16^0, 16^{-1}, . . . (expressed in the decimal system) where 16^0 place is unit's place.

Example 13.7

Convert $(A2F.D)_{16}$ into the decimal system.

Solution

$$\begin{aligned}(A2F.D)_{16} &= A \times 16^2 + 2 \times 16^1 + F \times 16^0 + D \times 16^{-1} \\ &= 10 \times 256 + 2 \times 16 + 15 \times 1 + 13\ (0.0625) \\ &= 2560 + 32 + 15 + 0.8125 \\ &= 2607.8125\end{aligned}$$

Example 13.8

Convert 2607.8125 into hexadecimal system.

Solution

$$
\begin{array}{llll}
16\,\underline{)\,2607} & & & \\
16\,\underline{)\;\;162} & \ldots & 15 & (F) \\
16\,\underline{)\;\;\;10} & \ldots & 2 & \\
\qquad\;\; 0 & \ldots & 10 & (A)
\end{array}
\Big\uparrow
\qquad
13\,(D)\;
\begin{array}{r}
0.8125 \\
\times\; 16 \\
\hline
13.0000 \\
\hline
\end{array}
$$

$$\therefore \qquad 2607.8125 = (A2F.\,D\,)_{16}$$

The conversions among binary, octal, decimal and hexadecimal systems play an important role in computers. Since $2^3 = 8$ and $2^4 = 16$ each octal digit can be made to correspond to a three binary digit block and each hexadecimal digit to a four binary digit block. The following examples illustrate this idea employed in the conversion from one numeral system to the other.

Example 13.9

Convert $(10100111001011)_2$ into (a) octal system and (b) hexadecimal system.

Solution

(a) Starting from the right end of the numeral, group all the digits (as shown below) into blocks of three, appending some zeros, if necessary, on the left of the given numeral. Thus, we have

$$\overline{010}\;\;\overline{100}\;\;\overline{111}\;\;\overline{001}\;\;\overline{011}$$

(Note that one zero is appended.)

Replace each block (considering it as a binary numeral) by the corresponding octal digit. So we get

$$\begin{array}{ccccc} \overline{010} & \overline{100} & \overline{111} & \overline{001} & \overline{011} \\ 2 & 4 & 7 & 1 & 3 \end{array}$$

$$(10100111001011)_2 = (24713)_8$$

Explanation

Why we group all the digits into blocks of three, will be obvious from the following:

$$
\begin{aligned}
010\;100\;111\;001\;011 &= 0\cdot 2^{14} + 1\cdot 2^{13} + 0\cdot 2^{12} + 1\cdot 2^{11} + 0\cdot 2^{10} + 0\cdot 2^{9} \\
&\quad + 1\cdot 2^{8} + 1\cdot 2^{7} + 1\cdot 2^{6} + 0\cdot 2^{5} + 0\cdot 2^{4} + 1\cdot 2^{3} + 0\cdot 2^{2} + 1\cdot 2^{1} + 1\cdot 2^{0} \\
&= [\,0\cdot 2^{2} + 1\cdot 2^{1} + 0\cdot 2^{0}\,]\,(2^3)^4 + [\,1\cdot 2^{2} + 0\cdot 2^{1} + 0\cdot 2^{0}\,]\,(2^3)^3 \\
&\quad + [1\cdot 2^{2} + 1\cdot 2^{1} + 1\cdot 2^{0}\,]\,(2^3)^2 + [\,0\cdot 2^{2} + 0\cdot 2^{1} + 1\cdot 2^{0}\,]\,(2^3)^1 \\
&\quad + [\,0\cdot 2^{2} + 1\cdot 2^{1} + 1\cdot 2^{0}\,]\,(2^3)^0 \\
&= [\,(010)_2\,]8^4 + [\,(100)_2\,]8^3 + [\,(111)_2\,]8^2 \\
&\quad + [\,(001)_2\,]8^1 + [\,(011)_2\,]8^0 \\
&= [\,(010)_2\;(100)_2\;(111)_2\;(001)_2\;(011)_2\,]_8 \\
&= [\,\overline{010}\;\;\overline{100}\;\;\overline{111}\;\;\overline{001}\;\;\overline{011}\,]_8 \\
&= (2\;4\;7\;1\;3)_8
\end{aligned}
$$

(b) Starting from the right end of the numeral, group each 4 digits into a block appending some zeros, if necessary, on the left end as shown. Then, replace each block by the corresponding hexadecimal digit.

$$\begin{array}{cccc} \overline{0010} & \overline{1001} & \overline{1100} & \overline{1011} \\ 2 & 9 & C & B \end{array}$$

$$\therefore \quad (10100111001011)_2 = (29CB)_{16}$$

Note: The reason for grouping all the digits into blocks of four, appending zeros at the left whenever necessary, is similar to the explanation given above. In the case of octal system we made blocks of 3 digits because $8 = 2^3$ and in the case of hexadecimal system we made blocks of 4 digits because $16 = 2^4$.

Example 13.10

Convert $(67025)_8$ into binary system.

Solution

Replace each digit of the octal numeral by the corresponding 3- digit binary numeral.

$$\begin{array}{ccccc} 6 & 7 & 0 & 2 & 5 \\ 110 & 111 & 000 & 010 & 101 \end{array}$$

$$\therefore \quad (67025)_8 = (110\ 111\ 000\ 010\ 101)_2$$

Example 13.11

Convert the hexadecimal numeral $2CF3A$ into the binary system.

Solution

Replace each digit of the given numeral by the corresponding 4- digit binary numeral.

$$\begin{array}{ccccc} 2 & C & F & 3 & A \\ 0010 & 1100 & 1111 & 0011 & 1010 \end{array}$$

$$\therefore \quad (2CF3A)_{16} = (10\ 1100\ 1111\ 0011\ 1010)_2$$

EXERCISE 13.1

1. Convert the following binary numerals into decimal form:
   (i) 100110011 (ii) 11011011011

(iii) 11100011001 (iv) 101010101010101

2. Convert the following decimal numerals into binary form:
(i) 182 (ii) 225 (iii) 1000 (iv) 1024 (v) 1262

3. Convert the following decimals into binary form:
(i) 12.125 (ii) 26.25 (iii) 39.625 (iv) 74.1875

4. Convert the following binary numerals into decimals:
(i) 101.101 (ii) 11.011 (iii) 1010.1011 (iv) 111.1111

5. Convert the following octal forms into decimal form:
(i) 236.2 (ii) 426.05 (iii) 324.11 (iv) 235.25

6. Convert the following decimal numerals into octal numerals:
(i) 146.5 (ii) 267.25 (iii) 438.125 (iv) 692.625

7. Convert the following hexadecimal forms into decimal forms:
(i) *B2F*.5 (ii) *ABC*.2 (iii) 1*C*3.*D* (iv) 2*AF*. *A*

8. Convert the following decimal numerals into hexadecimal system:
(i) 465.5 (ii) 1046.25 (iii) 3465.125 (iv) 4246.625

9. Convert the following binary forms into (i) octal form (ii) hexadecimal form:
(i) 1010111010 (ii) 11001101101
(iii)1100110110111 (iv) 10100111010011

10. Convert the following octal numerals into binary numerals:
(i) 247 (ii) 1527 (iii) 2064 (iv) 32321

11. Convert the following octal numerals into binary numerals:
(i) 2*BCD* (ii) *A*7 *D F* (iii) *A B C D* (iv) *E*1*B F* 2

13.5 Computer Arithmetic

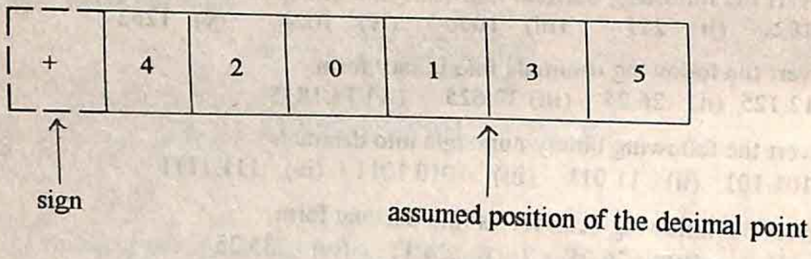
There are two types of arithmetic available in a computer. They are .

1. integer arithmetic and
2. real arithmetic or floating point arithmetic.

In mathematics 2 and 2.0 are the same. But in computer they are not considered to be the same. In computer terminology 2 is an integer and 2.0 is real number. If we instruct the computer to divide 7 by 2 (both being integers) it gives the answer 3 (the fractional part is lost). This is integer arithmetic. On the other hand if we instruct the computer to divide 7.0 by 2.0 (both being real) it gives the answer as 3.5. This is called real arithmetic.

Binary number system is used to store and manipulate information inside the computer. But for easier understanding of the computer arithmetic we shall assume in the following discussion that decimal numbers are stored and operated upon in the computer. For this purpose, let us assume that we have a (hypothetical) computer and its memory location can hold only 6 decimal digits with additional provision for the sign (+ or –). You may note here that each memory location (or a register) of any computer can hold only a finite number of digits.

One method of storing a real decimal number in our computer memory location is to fix the position of the decimal point and represent the number as shown in Fig. 13.4.



Memory Location
Fig. 13.4

The real decimal number stored as shown in Fig. 13.4 is

$$+\,4201.35$$

This representation is called *fixed point representation* since the position of the decimal point is assumed to be fixed after 4 positions from left. If we use fixed point representation as above, the largest positive number that we can store in our computer is 9999.99 and the smallest positive number is 0000.01. This range is quite inadequate for any computer problem solving. So we adopt a different method of representing real decimal numbers which is known as *floating point representation*.

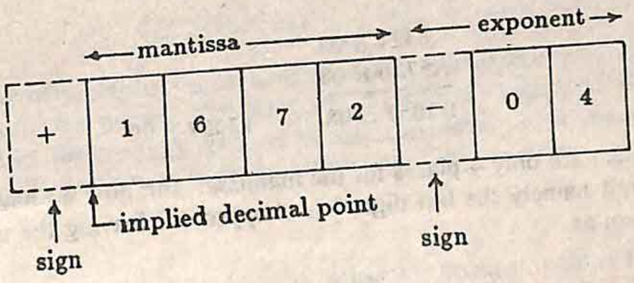
What is floating point representation ? Let us take an example. You can see that the number 123.768 can be writen in several forms, by adjusting the exponent, as shown below.

$$123.768 = 12.3768 \times 10^1 = 1.23768 \times 10^2 = 12376.8 \times 10^{-2}$$
$$= 0.0123768 \times 10^4 = 1237680 \times 10^{-4}, \text{ and so on.}$$

Since we can 'float' the decimal point within the numeral, all these representations are called floating point representations of the same number.

The floating point representations of a real number has two parts, namely (i) mantissa part and (ii) the exponent part. For example, in 12.3768×10^1, 12. 3768 is the mantissa and 1 is the exponent; in 0.0123768×10^4, 0.0123768 is the mantissa and 4 is the exponent; in 1237.68×10^{-1}, mantissa is 1237.68 and exponent is -1; in 123.768, mantissa is 123.768 and exponent is 0. As you have seen mantissa part and exponent part have their own signs.

For storing a real number in a computer memory location, we use that floating point representation of the number in which mantissa is between 0 and 1 and its first digit is non-zero. Such a floating point representation is called the *normalised floating point representation* of the number. For example $.1672 \times 10^{-4}$ is normalised floating point representation, while $.01672 \times 10^{-3}$ or 1.672×10^{-5} are not normalised floating point representations. The normalised floating point representation $.1672 \times 10^{-4}$ is written as $.1672\,E - 04$. The number to the left of E is the mantissa and the number to the right of E is the exponent. The number $.1672\,E - 04$ is stored in the computer memory location as shown in Fig. 13.5. Notice that the representation used in Fig. 13.5 is the normalised floating point representation.



Memory Location
Fig. 13.5

Every floating point representation can be expressed as normalised floating point representation by adjusting the exponent. For example

$$.0214\,E\ 06 = .0214 \times 10^{6} = .2140 \times 10^{5} = .2140\,E\ 05$$

$$3.264\,E\ 12 = 3.264 \times 10^{12} = .3264 \times 10^{13} = .3264\,E\ 13$$

$$26.752\,E - 04 = 26.752 \times 10^{-4} = .26752 \times 10^{-2} = .26752\,E - 02$$

As you can see, when we use normalised floating point representation, the range of numbers (in magnitude) that can be stored in our (hypothetical) computer will be $.1000 \times 10^{-99}$ to $.9999 \times 10^{-99}$, which is obviously much larger than the range in the case of fixed point representation.

Let us discuss how arithmetic operations are performed with normalised floating point representations.

Addition

The general principle for adding two normalised floating point representations is that their exponents should be made equal before adding them. The details are explained through the following examples.

Example13.12

Add $.2642\,E\ 05$ and $.3781\,E\ 05$.

Solution

Here the exponents are equal. So add the mantissas which gives the mantissa of the sum. The exponent will remain the same as the exponents of the summends. Thus

$$\begin{array}{rr} & .2642\,E\ 05 \\ + & \underline{.3781\,E\ 05} \\ & \underline{.6423\,E\ 05} \end{array}$$

Example 13.13

Add .6321 E 08 and .5736 E 08.

Solution

$$\begin{array}{r} .6321\ E\ 08 \\ +\ \ .5736\ E\ 08 \\ \hline 1.2057\ E\ 08 \\ \hline \end{array} = .12057\ E\ 09$$

In our computer there are only 4 places for the mantissa. But now we have 5 digits. So the least significant digit namely the last digit 7 is chopped off, leaving the mantissa as .1205. Thus, we get the sum as

$$.1205\ E\ 09$$

$$\therefore \quad .6321\ E\ 08 + .5736\ E\ 08 = .1205\ E\ 09$$

Example 13.14

Add .7643 E 04 and .5231 E 05.

Solution

Here the exponents are not the same. So we will retain the number with greater exponent as it is and change the other number suitably so that the exponents of the two numbers become equal.

$$\begin{array}{r} .7643\ E\ 04 \\ +\ \ .5231\ E\ 05 \\ \hline \\ \hline \end{array} \rightarrow \begin{array}{r} .0764\ E\ 05 \\ +\ \ .5231\ E\ 05 \\ \hline \\ \hline \end{array} \rightarrow \begin{array}{r} .0764\ E\ 05 \\ +\ \ .5231\ E\ 05 \\ \hline .5995\ E\ 05 \\ \hline \end{array}$$

Remark

The reason why we retained the number with greater exponent as it is and changed the other suitably is as follows. If we would have worked with the exponent that is smaller (namely 4 in this example) the other number .5231 E 05 would have required to be changed to 5.231 E 04. But mantissa of this being greater than 1, it can not be stored in our (hypothetical) computer. This is so because the implied position of the decimal point is such that there is no place for storing any digit to its left.

Example 13.15

Add .7132 E 99 to .5643 E 99.

Solution

$$\begin{array}{rr} & .5643\,E\,99 \\ + & .7132\,E\,99 \\ \hline & (1.2775\,E\,99) \\ \hline \end{array} \quad \rightarrow \quad .1277\,E\,100$$

Now our (hypothetical) computer cannot hold 100 in the exponent place as it has three digits. In other words the number that we got as the sum is larger than the largest number that can be stored in the memory location of our computer. This condition is called an overflow condition. The computer gives an indication of this condition, whenever it occurs.

Subtraction

Subtracting a number is nothing but adding its negative. So all principles for adding numbers hold for subtraction also. The following examples illustrate the procedure.

Example 13.16

Subtract .2734 E 05 from .6384 E 05.

Solution

We note that the exponents of the numbers are equal. Thus

$$\begin{array}{rr} & .6384\,E\,05 \\ - & .2734\,E\,05 \\ \hline & .3650\,E\,05 \\ \hline \end{array}$$

Example 13.17

Subtract .7216 E 05 from .2536 E 06.

Solution

We note that the exponents are unequal. As in the case of addition, we will retain the number with larger exponent and change the other number suitably. Thus

$$.7216\,E\,05 \quad = \quad .0721\,E\,06 \quad \text{and}$$

$$\begin{array}{rr} & .2536\,E\,06 \\ - & .7216\,E\,05 \\ \hline & \\ \hline \end{array} \quad \rightarrow \quad \begin{array}{rr} & .2536\,E\,06 \\ - & .0721\,E\,06 \\ \hline & .1815\,E\,06 \\ \hline \end{array}$$

Example 13.18

Subtract .4624 $E - 12$ from .4657 $E - 12$.

Solution

$$\begin{array}{r} .4657\,E-12 \\ -\ \ .4624\,E-12 \\ \hline .0033\,E-12 \\ \hline \end{array}$$

The answer is not in the normalised form. Therefore, it is to be converted into normalised form.

$$.0033\,E-12 = .3300\,E-14$$

$\therefore$ Answer is $.3300\,E-14$

Example 13.19

Subtract $.5436\,E-99$ from $.5570\,E-99$.

Solution

$$\begin{array}{r} .5570\,E-99 \\ -\ \ .5436\,E-99 \\ \hline .0134\,E-99 \\ \hline \end{array} \quad \rightarrow \quad .1340\,E-100$$

(The answer is expressed in the normalised form)

The exponent part of the answer has 3 digits and our computer has provision for it. In other words the answer we got is smaller than the smallest number that can be stored in the memory of our computer. This condition is known as *underflow condition*. The computer indicates this condition when it occurs.

Multiplication

You know

$$(a \times 10^{p}) \times (b \times 10^{q}) = ab \times 10^{p+q}$$

So, to multiply two numbers in normalised floating point representation (i) we multiply their mantissas to obtain the mantissa of the product and (ii) add the exponents to obtain the exponent of the product. The final answer is given in the normalised floating point form. The following examples illustrate the procedure.

Example 13.20

Multiply $.2642\,E\,09$ by $.4125\,E-04$.

Solution

$$.2642 \times .4125 = .1089\ \underbrace{8250}_{\text{discarded}}$$

Sum of the exponents $= 05$

$$\therefore \quad .2642\,E\,09 \times .4125\,E-04 = .1089\,E\,05$$

Example 13.21

Multiply .4673 E 62 by .3423 E 41.

Solution

In the resulting product the exponent would be $62 + 41 = 103$. There is no provision in our computer to store 3 digits in the exponent place. This signals the overflow condition.

Example 13.22

Multiply .3654 $E - 72$ by .7342 $E - 43$.

Solution

In the resulting product the exponent would be $-72 - 43 = -115$. This signals the underflow condition.

Division

You know

$$\frac{a \times 10^p}{b \times 10^q} = \left(\frac{a}{b}\right) 10^{p-q}, \quad (b \neq 0)$$

So, to divide one number (first number) by another number (second number) (i) divide the mantissa of the first number by the mantissa of the second number to obtain the mantissa of the quotient and (ii) subtract the exponent of the second from the exponent of the first to obtain the exponent of the quotient. Finally the quotient is given using the normalised floating point representation.

Example 13.23

Divide .4267 E 15 by .2437 $E - 02$.

Solution

$$\frac{.4267}{.2437} = 1.7509\,\underbrace{23266} = 1.7509$$

$$\frac{.4267\,E\,15}{.2437\,E-02} = 1.7509\,E\,(15+2)$$

$$= 1.7509\,E\,17 = .1750\,E\,18$$

Subtraction in Integer Arithmetic

How subtraction is carried out in integer arithmetic is interesting to note. To subtract b from

a, the computer finds the complement of b, adds it to a and makes some adjustments in the result obtained. We will see how this is done through some examples.

By direct subtraction, we have

$$\begin{array}{r} 43012613 \\ -\quad 738047 \\ \hline 42274566 \\ \hline \end{array}$$

Let us understand what is meant by *complement of a digit*. There are two types of complements of a digit: the 9's complement and the 10's complement. The 9's complement of 2 is 9 − 2, i.e. 7. The 9's complement of 4, 0, 6, are 5, 9, 3 respectively. The 9's complement of a number is obtained by replacing each digit of the numeral by its 9's complement. For example, the 9's complement of 72601 is 27398.

We have seen above that when we subtract 738047 from 43012613 we get 42274566. Let us see what we get when we add the 9's complement of 738047 to 43012613. In this process, we will first make the number of digits in the minuend and subtrahend to be equal by writing zeros to the left of the subtrahend. In this example the minuend has 8 digits and the subtrahend has 6 digits. So we write the subtrahend as 00738047. The 9's complement of 00738047 is 99261952. Let us add this 9's complement to 43012613.

$$\begin{array}{r} 43012613 \\ +\quad 99261952 \\ \hline 142274565 \\ \hline \end{array}$$

In the result there is an extra digit namely 1 at the extreme left. Remove this digit from that place and add to the number left out. We get 42274565 + 1, i.e. 42274566, the same number that we got as the result of direct subtraction.

Direct Subtraction

$$\begin{array}{r} 43012613 \\ -\quad 00738047 \\ \hline 42274566 \leftarrow \\ \hline \end{array}$$

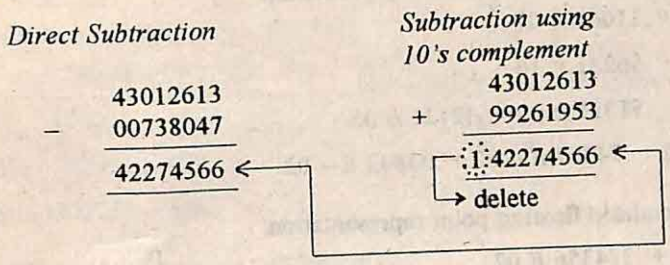
Subtraction using 9's complement

$$\begin{array}{rl} 43012613 & \\ +\quad 99261952 & \text{(9's complement)} \\ \hline \boxed{1}\,42274565 & \\ \rightarrow\; +\,1 & \\ \hline \rightarrow\; 42274566 & \\ \hline \end{array}$$

The 10's complement of a number is obtained by finding the 9's complement and adding 1 to it. The 10's complement of 00738047 is 99261952 + 1 i.e., 99261953. Let us see what we get by adding this 10's complement to 43012613.

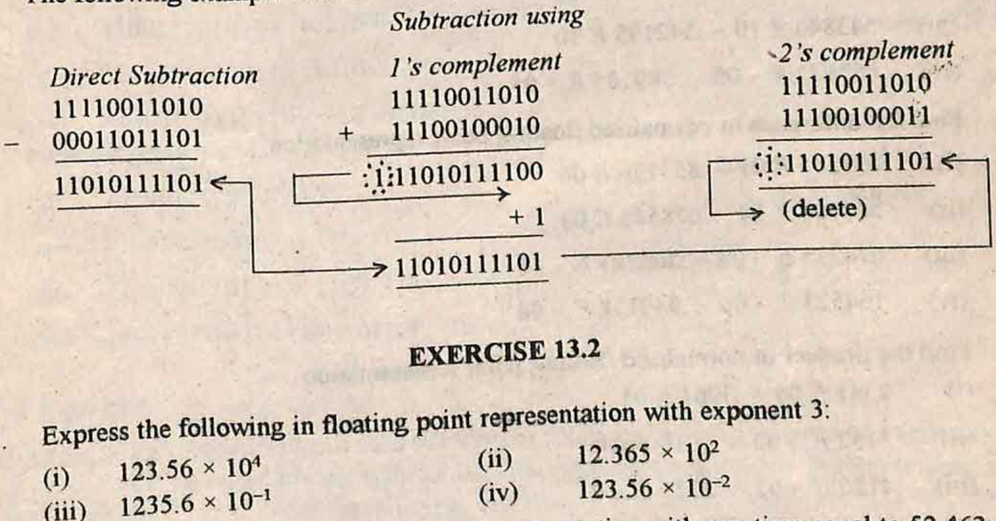
$$\begin{array}{r} 43012613 \\ +\quad 99261953 \\ \hline 142274566 \\ \hline \end{array}.$$

Delete the extra digit 1 on the extreme left. You get 42274566, which is the same number that we got as the result of direct subtraction.



The process of subtraction using addition of complement as explained above holds in all base systems. In the case of binary system, we have to take the 1's complement or 2's complement. The 1's complements of the binary digits 0 and 1 are 1 and 0 respectively. So the 1's complement of a binary numeral is simply obtained by changing the zeros into ones and ones into zeros. It is very easy to implement this conversion in a computer. That is why subtraction is carried out in a computer using addition of 1's complement or 2's complement. The 2's complement is obtained by adding 1 to the 1's complement.

The following example illustrates the process of subtraction using 1's / 2's complements.



EXERCISE 13.2

1. Express the following in floating point representation with exponent 3:
   (i) 123.56×10^4 (ii) 12.365×10^2
   (iii) 1235.6×10^{-1} (iv) 123.56×10^{-2}
2. Express the following in floating point representation with mantissa equal to 52.463:
   (i) 5246.3×10^2 (ii) 5.2463×10^{-2}
   (iii) 52.463×10^{-2} (iv) $.52463 \times 10^{-3}$
3. Express the following in normalised form of floating point representation:
   (i) 1.24365×10^6 (ii) 12.4365×10^6

(iii) $.0124365 \times 10^6$ (iv) $.000124365 \times 10^6$

4. Find the sum in normalised floating point representation.
   (i) $.23452\ E\ 07 + .31065\ E\ 07$
   (ii) $.74315\ E\ 10 + .56231\ E\ 10$
   (iii) $.41362\ E\ 05 + .51321\ E\ 05 + .42121\ E\ 05$
   (iv) $.38756\ E - 02 + .74387\ E - 02 + .63843\ E - 02$

5. Find the sum in normalised floating point representation.
   (i) $.436527\ E\ 05 + .274356\ E\ 07$
   (ii) $.623543\ E\ 04 + .578132\ E\ 05$
   (iii) $.246712\ E\ 04 + .147238\ E\ 03 + .413895\ E\ 05$
   (iv) $.467342\ E - 01 + .963856\ E\ 01 + .453218\ E\ 01$

6. Find the difference in normalised floating point representation.
   (i) $.642752\ E - 03 - .374804\ E - 03$
   (ii) $.824631\ E\ 06 - .743215\ E\ 06$
   (iii) $.543846\ E\ 10 - .542195\ E\ 10$
   (iv) $.653172\ E - 05 - .589185\ E - 05$

7. Find the difference in normalised floating point representation.
   (i) $.471255\ E\ 07 - .853196\ E\ 06$
   (ii) $.583904\ E\ 10 - .678542\ E\ 08$
   (iii) $.674235\ E - 04 - .265281\ E - 05$
   (iv) $.164523\ E - 06 - .849138\ E - 08$

8. Find the product in normalised floating point representation.
   (i) $.2341\ E\ 05 \times .3061\ E\ 01$
   (ii) $.3152\ E - 02 \times .1010\ E\ 04$
   (iii) $.4125\ E - 02 \times .2121\ E - 03$
   (iv) $.5125\ E - 01 \times .4021\ E\ 02$

9. Divide
   (i) $.6712\ E\ 10$ by $.2643\ E\ 04$
   (ii) $.4396\ E\ 05$ by $.3512\ E - 02$

(iii) $.6123\,E-07$ by $.2132\,E-03$

(iv) $.8642\,E\,02$ by $.2562\,E\,02$

10. Subtract using the 9's complement.
    (i) 4675214 – 2134527
    (ii) 3529723 – 496713
    (iii) 674235 – 58231
    (iv) 4501952 – 8567
11. Subtract using the ten's complement.
    (i) 374813 – 143956
    (ii) 5439852 – 26839
    (iii) 758324 – 64925
    (iv) 4936852 – 49245
12. Subtract using 1's complement.
    (i) 10110011101 – 11101011
    (ii) 110011101010 – 10110011
    (iii) 1011001100011 – 1010110111
    (iv) 110011011101101 – 1010110011
13. Subtract using 2's complement.
    (i) 10100111001 – 1101011
    (ii) 1100110100101 – 110011011
    (iii) 1111001001001 – 11001111
    (iv) 110110110110111 – 1011001101

13.6 Algorithm

Computer, as everybody knows, is used to carry out computations. As we generally understand, a computation involves additions, subtractions, multiplications, divisions, finding powers and roots of numbers, etc. These are all what are known as numeric computations, i.e. those involving numbers. In the context of computers, the word computation is to be understood in a broader or extended sense. A computer can be used to sort out a given list of names of persons in an alphabetical order. This type of work is called a non-numeric computation. In early years after the computers came into being, they were predominantly used for numeric computations associated with solving equations, etc. Over the years since then, the mix of the type of work computers do has changed so much. Nowadays computers are employed more for non-numeric computation.

How should we get a computation — numeric or non-numeric — done by computer ? We have seen in earlier classes that whether a problem is small or big, simple or complicated, we think about it and come up with a systematic step-by-step procedure which, if followed meticulously, will lead us to solution of the problem. A technical term for a step-by-step procedure is *algorithm*. In development of algorithm, sequence, selection, and repetition (or iteration) play an important role . We shall study these aspects of an algorithm now.

Sequence

Suppose that we want to find the value of the expression $a^3 + 4ab + b^2$ for given values of a and b. Algorithm (i.e. step by step procedure) for achieving this will consist of steps given in Fig. 13.6 to be carried out one after the other.

1. get the value of a
2. get the value of b
3. Calculate a^3, call it S
4. Calculate $4ab$, call it T
5. Calculate b^2, call it V
6. Find the sum $S + T + V$, call it M
7. Write the value of M as answer.

Steps of an algorithm to evaluate $a^3 + 4ab + b^2$,
given the values of a, b

Fig. 13.6

This algorithm, you will agree, is very straightforward, consisting of simple steps which are to be carried out one after the other. We say that such an algorithm is a *sequence* of steps, meaning that

(a) At a time only one step of the algorithm is to be carried out.

(b) Every step of the algorithm is to be carried out once and only once; none is repeated and none is omitted.

(c) The order of carrying out the steps of the algorithm is the same as that in which they are written.

(d) Termination of the last step of the algorithm indicates the end of the algorithm.

Here afterwards we shall follow the convention that (i) the successive steps in a sequence will be written on successive lines and hence (ii) steps will not be necessarily numbered as they are in Fig. 13.6.

Selection

It is easy to appreciate that an algorithm which is completely composed of only a sequence, will not be sufficient to solve any type of problem. For example you know the definition of $|x|$ when x is a real number. The definition is

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

Now suppose that we want to write an algorithm to obtain the value of $|x|$, knowing x. It will be as in Fig. 13.7

```
1.  get the value of $x$
2.  if $x \geq 0$
        then   required value is $x$
        else   required value is $-x$
3.  write the required value obtained.
```

Algorithm to obtain value of $|x|$, x real

Fig. 13.7

Thus, we have provided the person or computer that will execute the algorithm with an ability to choose the step to be carried out depending on the circumstance of x being either non-negative or negative. This ability is called *selection* . It is extremely useful for writing non-trivial meaningful algorithms. The power of selection is that it permits that different paths could be followed, depending on the requirement of the problem, by the one who executes the algorithm.

In Fig. 13.7, selection is expressed by using the special words **if, then, else**. Further all that is written using these special words constitutes one step; viz step 2 in Fig. 13.7. Note the way it is written without anything appearing below the word **if** till that step is over. This is known as *indentation*. The words **if, then**, and **else** being special are underlined while writing by hand and are written in bold face (in printing). The words **then** and **else** come with exactly same indentation with respect to word **if**.

The algorithm in Fig. 13.7 is a *sequence* of 3 steps; the second step is of *selection* type. Any algorithm will be a sequence of steps; some of its steps may be of *selection* type. You think over and you will realize that it is impossible to write an algorithm of any significant practical use without having steps of selection type. At the same time, just sequence and selection is not sufficient. .We show this in an example in next section.

Iteration or Repetition

Example 13.24

Write an algorithm to find the first prime number greater than a given positive integer.

Solution

A little thought will tell you that a possible step-by-step procedure to achieve our objective will be one given in Fig. 13.8.

```
consider the given integer
add 1 to it
test new number for primeness          }
if it is prime                         }  S
  then write it down                   }
  else add 1 to it
       test new number for primeness          }
       if it is prime                         }  S
         then write it down                   }
         else add 1 to it
              test new number for primeness          }
              if . . .                               }  S
              . . .                                  }
```

Algorithm for finding the first prime greater than a given positive integer

Fig. 13.8

This shows that we need to repeat certain steps before algorithm terminates after giving an answer. This is technically known as *iteration* or *repetition*. The way of writing adopted in Fig. 13.8 presents a difficulty. It is not at all clear as to how many times we should write the same set S (indicated at right in Fig. 13.8) of instructions so that the whole collection of instructions will give the required answer. The number of times S has got to be repeated will depend on the given number. The difficulty faced here is resolved by introducing a way of writing iteration in algorithms. The algorithm in Fig. 13.8 is written as shown in Fig. 13.9.

```
consider the given number
repeat
        add 1 to it
until   the new number is prime
write   the new number

              or

consider the given number
add 1 to it
while   the new number is not prime
        do add 1 to it
write   the new number
```

Two different ways of writing iteration occurring in Fig. 13.8

Fig. 13.9

Just as selection in an algorithm was expressed using the words **if**, **then** and **else**, the iteration (i.e. repetition) in an algorithm is expressed using either the pair of words **repeat** and **until** or the pair of words **while** and **do**. Note the indentation used while using any of these pair of words. Study carefully the difference between the uses of **repeat-until** pair and **while-do** pair. We shall talk more about them a little later.

Remarks

We have taken very simple examples. Algorithms written above are straightforward and easily understood. Main purpose there has been to show how we need these three aspects, viz: sequence of steps, selection of steps and repetition of steps in any worthwhile algorithm. In fact, we do not need anything else than these 3 basic operations for any algorithm, howsoever complicated or difficult. Bohm and Jacopini have proved a very profound theoretical result that if an algorithm exists to solve a problem, then it can always be expressed using only **if-then-else** construct, **while-do** construct and a sequence of steps. Even though these 3 constructs are sufficient theoretically to develop any algorithm, we shall use **repeat-until** construct (introduced above) and a **for** construct (to be introduced later) also to make the task of writing an algorithm more straightforward, easier and pleasant.

In our algorithm we shall also use the word **comment**. Whatever follows after this word **comment** in that step of an algorithm is only by way of explanation. This is meant to facilitate understanding of the algorithm while reading it. While executing the algorithm this comment statement has no effect or plays no role and is ignored. You are encouraged to write these comment statements to make the algorithm readable. But be brief in what is written after the word **comment**.

13.7 Variable Name or Identifier

You are familiar with the concept of a variable. You have been using single letters (capital or small) like x, y, X, Y to denote a variable. By combining the variables with operations we get expressions like $x^2 + 2x - 9$, $XY + 6$. When we use symbols like x or Y to denote a variable, we shall alternatively say that x is the name of one variable, Y is the name of another variable and so on. Thus, we shall call x, Y, X as *variable names*. We also use meaningful variable names. For example, in the formula

$$I = PTR \qquad \ldots (13.1)$$

I stands for Interest, P for Principal, T for Time in years and R for Rate of interest. Thus instead of writing the formula as $w = xyz$ and saying that w, x, y, z respectively stand for interest, principal, time and rate of interest, we use the formula (13.1) in which first letters of the words Interest, Principal, Time and Rate are chosen to indicate the variable names. Such variable names are easy to remember and are known as *mnemonic* variable names. Incidentally, recall from Chapter 19 from Class XI book that while writing algorithms for computers we write formula (13.1) as

$$I = P * T * R$$

using * for multiplication sign which is explicitly written every time.

A variable name identifies a certain variable. For example , in formula (13.1), variable name I identifies the interest, etc. Hence, the term *identifier* is used synonymously with the term variable name.

In mathematics, generally single letters like x, y, I, P, T are used as variable names. But sometimes in order to convey a given situation in a better way, it has been the practice to use more than one letter for a single variable name. For example, you know that Δx is used as a variable name to denote the increment in the value of x. Especially in the area of computing it is found convenient and, hence, it has become customary to use more letters to denote a variable name. For example, we shall prefer to write formula (13.1) as

$$\text{INT} = \text{PRIN} * \text{TIME} * \text{RATE} \qquad \ldots (13.2)$$

where variable name INT is used for interest and so on for other variables in the formula. One can use INTEREST as well instead of INT. You will agree that form (13.2) brings out more vividly the variable names than form (13.1). As another example, instead of just S, it would be better to use 'speed' or 'spd' as variable names for the variable speed. Such choice of variable names convey, by sight itself, the variables for which these names stand. Hence, such variable names are called mnemonic variable names as mentioned earlier. NUM, DEN can be used as mnemonic variable names for numerator and denominator respectively. We advice you to use meaningful mnemonic variable names. Dictionary meaning of "mnemonic" is "easy to commit to memory".

13.8 Assignment Operation

While writing algorithms or flowcharts, many times we are required to assign a value to a variable. For example, see the step 3 in Fig. 13.6. There we say "calculate a^3 and call it S". This means the variable S is assigned the value a^3. We shall write this symbolically as

$$S \leftarrow a^3 \qquad \ldots (13.3)$$

Similarly, we may like to give value 1 to variable N in the beginning to start keeping count. We shall write it as

$$N \leftarrow 1 \qquad \ldots (13.4)$$

If we want to give value 20 to a variable x, we shall write this instruction as

$$x \leftarrow 20 \qquad \ldots (13.5)$$

Thus, *assignment is an operation* and we use ← as an *assignment symbol.*

'$x \leftarrow 20$' is read as 'x becomes 20' meaning thereby that the value of variable x becomes 20 at that stage of the algorithm or flowchart at which this statement is written. When this statement or instruction is executed by a computer (i) it chooses one memory location, (ii) puts 20 in that memory location, and (iii) that memory location is given a symbolic address x. That memory location always has its permanent numerical address. When this instruction is carried out that permanent numerical address is symbolically identified as x and is referred to as such by x in later portion of the algorithm or flowchart. After the execution of the instruction $x \leftarrow 20$, the contents of this memory location (with symbolic address x) becomes 20.

We shall always (i) write an arrow with its head pointing towards left, (ii) write the variable name on the left side of arrow head, and (iii) write the value assigned to the variable on the right side of the tail of the arrow.

Very important rule to be remembered and followed while using assignment symbol ← is the following: Only a single variable name should appear on the left side of the head of the arrow; no expression involving addition, multiplication etc. should appear there. Thus, writing

$$x^2 + x \leftarrow 20 \qquad \ldots (13.6)$$

is not allowed and will be considered as wrong and meaningless usage. On the other hand one can have an expression involving addition, multiplication etc. on the right side of the tail of the arrow. Thus, we can write

$$y \leftarrow 10 - 2 \qquad \ldots (13.7)$$

In such a case, the expression on the right hand side is evaluated and its value is assigned to the variable name appearing on the left. So in (13.7), the variable y is assigned the value 8.

A statement [such as (13.3), (13.4), (13.5), (13.7) but not (13.6)] which uses assignment symbol ← following above-mentioned rules is called an *assignment statement.*

Suppose we have the following as two given consecutive assignment statements in an algorithm

$$N \leftarrow 3$$
$$M \leftarrow N^2 + N - 1$$

You can see that both these are perfectly valid assignment statements as per rules laid down above. How are these statements executed ? First statement assigns (gives) the value 3 to the variable N. With this value assigned, the second statement can be executed. First, right hand side $N^2 + N - 1$ is evaluated (with N being 3 now) to be $3^2 + 3 - 1 = 11$ and M is assigned this value 11.

13.9 Assignment Symbol to Keep Count

While giving a method of solution to a problem, we may not know beforehand, as to how many times a certain action is required to be performed. (We have come across such a situation in the algorithm given in Figs. 13.8 and 13.9.) But we may like, or sometimes need, to keep count of the number of times that action is repeated. Value of the variable, used to keep this count, could be zero to start with and its value can be made to increase by 1 each time we repeat that action. Suppose N is the variable name that we use to keep this count. Then we use the assignment statement

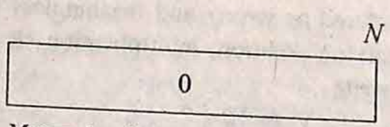
$$N \leftarrow 0 \qquad \ldots (13.8)$$

to set the count at zero to start with. Variable name N will continue to indicate the count even when its value gets changed by increment. The new value of N will become (old value of $N + 1$) every time we repeat the action. This can be written as

$$N \leftarrow N + 1 \qquad \ldots (13.9)$$

and read as 'N becomes $N + 1$'. As we first evaluate the right hand expression in an assignment statement, (13.9) indicates that '1 is added to current value of N and that is taken as new value of N'.

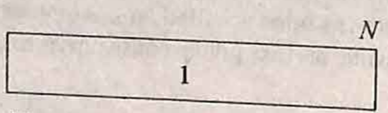
Remember that variable name N has a memory location associated with it. The symbolic address of that location is N. The content of that location is zero when (13.8) is executed. When (13.9) is executed first time after (13.8) contents of that location becomes 1. So value zero of N put earlier in that memory location is erased and 1 is put in it. Pictorially when (13.8) is executed we have



Memory location with symbolic address N after execution of $N \leftarrow 0$

Fig. 13.10

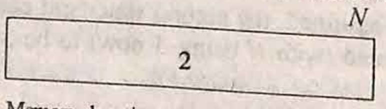
When (13.9) is executed first time, same memory location will look as



Memory location with symbolic address N after execution of $N \leftarrow N + 1$ first time

Fig. 13.11

When (13.9) is executed second time, same memory location will look as



Memory location with symbolic address N after execution of (13.9) second time

Fig. 13.12

Remarks

Some of you may feel that we can write assignment statement as '$x = 20$' rather than '$x \leftarrow 20$'. But then, when we want to increment the current value of N by, say 1, we would be required to write $N = N + 1$ which is absurd mathematically as it implies '$1 = 0$'. So it is advisable not to use symbol '=' as assignment symbol. Use of ← is, therefore, introduced.

It may be mentioned here that in the programming language FORTRAN '=' is used as an assignment symbol. '=' is a standard universal symbol for equality. So a new symbol for equality had to be invented for users of FORTRAN. It is unfortunate to use an existing, standard, well established and such a well-known symbol for equality for something different and then try to invent a new symbol for equality. This has been corrected in newer and better programming languages by keeping '=' for equality and using different symbol for assignment. In this light, we shall use ← only as an assignment symbol.

13.10 A Pseudo Language

The languages used by human beings for talking and writing among themselves are called *natural languages*. Thus English, Hindi, Tamil, German, Swahili etc. are examples of natural languages. Expression in a natural language can be ambiguous. You can appreciate this by seeing that different meanings can be attached to the statement "They are flying planes". In one meaning the word 'they' refers to the objects in the sky and 'flying' is an adjective of 'planes'. In that case the statement conveys certain objects are planes that are flying (and not grounded planes). In another meaning the word 'they' refers to persons and 'flying' is a present continuous of verb 'fly'. In that case the statement can be taken as an answer to a question "What are these persons doing ?"

Computer, being a machine, requires that there should be no ambiguity at all when we give instructions to it. Therefore, it is required that we express our algorithms in precise and unambiguous language. Languages used to communicate with a computer are known as *programming languages* as contrast to natural languages meant for communication among humans. Every computer has its own language known as machine language. You must have heard the names such as FORTRAN, PASCAL, COBOL. These are called high level programming languages. We do not want here to study a particular programming language as we do not want to get bound by its limitations and drawbacks. Our aim is to be able to write the algorithms to solve different problems. We shall like to express an algorithm in a language which is somewhat like English (as we are doing our study of this subject using English) and uses symbols and constructions which are characteristics of a good programming language.

We shall use meaningful mnemonic variable names (discussed in Section 13.7), assignment symbol ← (discussed in Sections 13.8 and 13.9), constructions employing **if-then-else**, **repeat-until**, **while-do** (introduced in Section 13.6) and other constructions employing word **for** (that will be introduced later) for writing an algorithm. We shall also require instructions to input data in an algorithm as well as instructions to output computed results from an algorithm. We shall introduce these below in this section. All these will constitute our language to present any algorithm. This language will not resemble in toto with any actual existing programming language but will have desirable characteristics of a good programming language (and hence we are using it). We shall call it a *pseudo language* .

Input- Output Instructions

We shall use the word **get** to have an instruction to fetch values of the variables. Thus

get x, y

in an algorithm means the values of the variables x and y are obtained and are available, for the remaining part of the algorithm, to operate upon. This further means that the numerical value of the variable x is put in a memory location and that memory location is given the symbolic name x , and similarly for variable y.

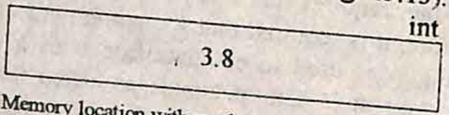
get name

in an algorithm means the value of the variable name 'name' is obtained and will be available to the remaining portion of the algorithm. Value of the variable name 'name' could be 'RAM' or 'GEETA' etc. (if we are dealing with the names of persons) or 'DELHI',' MADRAS' etc. (if we are dealing with cities). Remember that we are using meaningful variable names; otherwise variable name 'name' could be used for a numerical variable which can assume a value like 4 or 20.7.

Suppose a variable 'int' has value 3.8 The instruction

get int

will obtain for algorithm the value 3.8 of the variable int, put it in a memory location and give that memory location a symbolic name int (See Fig. 13.13).



Memory location with symbolic name and value in it after the execution of instruction get int

Fig. 13.13

Similarly, we shall use the word **output** to have an instruction to obtain the results worked out by an algorithm. Thus, the instruction

output y

means the value of y (in other words contents of the memory location with symbolic address y) will be obtained as an answer or result. When we work with a computer, effect of this instruction will be that the value of y (i.e. contents of memory location with symbolic address y) will be printed or displayed on screen depending on which output medium is used.

Whatever is put in quote marks in front of the word **output** in the instruction is put out (printed or displayed) verbatim. For example, suppose value of variable p is 37. Then the instruction

output 'answer' $= p$

when executed, wiill put out (print or display)

answer = 37

Algorithms of Figs. 13.6, 13.7, 13.9, when written in pseudo language will look like those presented in Figs. 13.14, 13.15, 13.16 respectively.

```
get $a, b$
$S \leftarrow a^3$
$T \leftarrow 4ab$
$V \leftarrow b^2$
$M \leftarrow S + T + V$
output $M$
```

Algorithm of Fig. 13.6 written in pseudo language

Fig. 13.14

```
get $x$
If $x \geq 0$
   then required value $\leftarrow x$
   else required value $\leftarrow -x$
output required value
```

Algorithm of Fig. 13.7 written in pseudo language
Fig. 13.15

```
get $n$
comment $n$ is the given positive integer
repeat
          $n \leftarrow n + 1$
          test $n$ for primeness
until     $n$ is prime
output    $n$
                 or

get $n$
comment $n$ is the given positive integer
$n \leftarrow n + 1$
while  $n$ is not prime
       do $n \leftarrow n + 1$
output $n$
```

Algorithm of Fig. 13.9 written in pseudo language
Fig. 13.16

EXERCISE 13.3

1. Improve the algorithm of Fig. 13.16 to find the first prime greater than a given positive integer. (Hint: All primes except 2 are odd numbers. Then why increase n by 1 always ?)
2. Write an algorithm to find first prime greater than a given integer, which can be positive, negative or zero. (Hint: First prime number is 2.)
3. What does the following algorithm do ?

```
get $x$
if $x \geq 0$
   then output "sq. root will be real"
   else output "sq. root will be imaginary"
```

4. Write an algorithm to find profit or loss given selling price, cost price and number of items sold.

5. Write an algorithm to declare whether the roots of $ax^2 + bx + c = 0, a \neq 0$ will be real or complex.

6. A is given number greater than 1. We want to find its square root within the accuracy of $\in$ by method of successive bisection. $\frac{A}{2}$ is taken as starting approximation. Complete the following algorithm to achieve our goal.

```
get A, $\in$
LEFT $\leftarrow 0$
RIGHT $\leftarrow A$
$x \leftarrow \frac{(\text{LEFT}+\text{RIGHT})}{2}$
while $|x^2 - A| > \in$
. . . . . . . .
```

7. Complete the following algorithm to add two matrices A and B.

```
get A, B
comment size of A, B is $m \times n$
for $i = 1$ to $m$
    do for $j = 1$ to $n$
        do . . . . . . .
        . . . . . . . . .
```

8. Write an algorithm to find $n\,!$ for given n.

9. Write an algorithm to find ${}^nC_r$ using the formula

$$ {}^nC_r = \frac{n\,!}{r!(n-r)!} $$

10. Write an algorithm to find ${}^nC_r$ using the formula

$$ {}^nC_r = \frac{n\,(n-1)\,(n-2)\ldots(n-r+1)}{r!} $$

13.11 Formal Explanation of Constructs

We have already seen the use of **if-then-else, repeat-until** and **while-do** constructs and have understood them intuitively. That is good in the beginning. Now here we shall give little more formal explanation of them, discuss some of their other aspects and introduce another construct, viz. **for** construct.

If-Then-Else Construct

We have seen in the algorithm given in Fig. 13.7, the use of this construct. The general form of this construct used to provide selection of actions is

```
if condition
  then step 1
  else step 2
```

When an instruction using this type of construct is executed, *condition* determines which of step 1 and step 2 is to be executed. If *condition* is true, step 1 is executed; otherwise (i. e. if *condition* is not true) step 2 is executed. This interpretation is obvious from the form of the construct.

A special case of this construct does not have the word else and as such has the general form:

```
if condition
   then step . . .
```

Obviously, when this form is used no action is taken when *condition* is false, and *step* is executed when *condition* is true. We need this form sometimes. We shall have examples of this situation in the sequel. Thus

```
if condition
   then step
```

and

```
if condition
   then step
   else do nothing
```

are equivalent constructs.

Example 13.24

Write an algorithm to decide whether the square root of a given real number will be real or imaginary.

Solution

```
get x
comment x is the given real number
if x ≥ 0
   then output " sq. root will be real "
   else output "sq. root will be imaginary "
```

Repeat-Until Construct

We used this construct in the algorithm given in Fig. 13.9. The general form of this construct, used when repetition of certain actions is required, is

```
repeat
          portion of algorithm
until     condition
```

When an instruction using this type of construct is executed, the portion of the algorithm (call it S) appearing between the words **repeat** and **until** is to be executed again and again until the *condition* mentioned after the word **until** is true. First S is executed; then the condition is tested. If the condition is true, the execution goes ahead with the algorithm appearing after the word **until**; if the *condition* is false, again S is executed and then the *condition* is again tested, and so on. Every execution of S modifies some variables in the algorithm and eventually after some repetitions, the *condition* becomes true. Thus, execution of the construct **repeat-until** is completed and the execution proceeds to the portion appearing after the word **until**.

Example 13.25

A number k is known to be present in a finite sequence A of numbers. A (1), A (2),. . . are the successive members of the given sequence A. Write an algorithm to find which member of A equals k.

Solution

```
get $A$
$n \leftarrow 0$
comment $n$ keeps count
repeat
        $n \leftarrow n + 1$
until   $A(n) = k$
output  "$k$ occurs as member number ", $n$ " of sequence $A$ "
```

While-Do Construct

We saw the use of this construct, in the algorithm of Fig. 13.9, as an alternative to the **repeat-until** construct. The general form of this construct, which is also used to provide repetition of instrution, is

```
while   condition
        do $T$
```

where T is a set of instructions. When this construct is executed, *condition* is evaluated first. If the *condition* is true, the set T of instructions is executed, and then the *condition* is evaluated again and so on; if the *condition* is false, execution of T is skipped and the execution of algorithm proceeds with the portion that would appear after T. Thus, condition is tested again and again till it is false. Every execution of T modifies some variables in the algorithm and eventually after some repetitions, the condition becomes false. This completes the execution of the **while-do** construct and the execution proceeds to the portion appearing after T.

You can notice that when **while-do** construct is used, T may not get executed even once. This happens when the condition is found not true at its first evaluation itself. On the other hand, in the use of **repeat-until** construct, the portion of the algorithm between **repeat** and **until** will be executed at least once. Second point worth noticing is that the execution of **while-do** construct is over when the condition therein is found false; on the other hand the execution of **repeat-until** construct is over when the condition therein is found true. Thus, these two constructs have duality in their form and nature. Thus

$$\left\{\begin{array}{l}\textbf{repeat}\\ \qquad S\\ \textbf{until}\quad B\end{array}\right\} \quad \text{and} \quad \left\{\begin{array}{l}S\\ \textbf{while} \sim B\\ \quad \textbf{do } S\end{array}\right\}$$

are equivalent. $\sim B$ stands for negative of B.

Example 13.26

Write an algorithm to go on halving a given number x until it becomes less than 1. (*Hint*: Note that the given number x may itself be less than 1 in which case nothing is expected to be done.)

Solution

```
get $x$
while $x \geq 1$
    do $x \leftarrow \frac{x}{2}$
```

For Construct

We have seen that **repeat-until and while-do** constructs are useful whenever iteration or repetition of steps is to be introduced in an algorithm. At the same time in using these constructs one need not be aware beforehand as to how many times a certain portion of the algorithm is to be repeated. For example, notice in Example 13.25, that the number of repetitions will depend on which member of squence equals k and will not be known until after the execution of complete algorithm. Similarly, in Example 13.26, how many times x will be required to be halved will depend how small or big x is to start with. In some practical problems, however, we may know, at the beginning of the algorithm itself, the number of times a certain portion of instructions is to be repeated. For example, suppose we want to add first ten positive even integers. This means we want to obtain the sum

$$2 + 4 + 6 + 8 + 10 + 12 + 14 + 16 + 18 + 20.$$

The algorithm for this can be written, using **repeat-until** construct, as follows:

```
sum ← 0
I ← 1
comment variable sum is initialized to zero and I keeps count
repeat
        sum ← sum +2 * I
        I ← I +1
until  I ≥ 11
```

The algorithm can also be written using **while-do** construct, as follows:

```
sum ← 0
I ← 1
while I ≤ 11
      do      sum ← sum + 2 * I
              I ← I + 1
```

For such situations in which number of repetitions to be carried out is known beforehand another construct known as **for-construct** is also available. Using it, above algorithm is written as

```
sum ← 0
for  I = 2 to 20 by 2
     do sum ← sum + I
```

The general form of **for-construct** is

```
for identifier = initial value to test value by increment
    do S
```

The words **for**, **to**, **by** and **do** are reserved words for this construct and they are underlined or written in bold face following the already stated convention. Initial value gives the starting value that the identifier should take when the S is executed; increment gives the value by which the value of identifier be increased after each execution of S; test value gives the value beyond which the value of the identifier should not go. Test value is the upper limit for the value of the identifier. Thus, when such a for-construct is executed, S gets executed repeatedly with the identifier starting with initial value and getting increased by increment after each execution of S and this repetition going on till identifier does not exceed the test value. One who writes the algorithm has to take care in fixing the initial value, test value and the increment for the identifier so that repetition does not become infinite. For example, instruction

```
for $I$ = 2 to 30 by − 3
    do $S$
```

will not do, but the instruction

```
for $I$ = 2 to − 12 by − 3
    do $S$
```

is vaild because identifier will take the values 2, −1, −4, −7 and −10 only.

When the value of increment is 1, it is customary not to mention it. Thus , the instructions

```
for $J$ = 3 to 18 by 1
    do $S$
```

and

```
for $J$ = 3 to 18
    do $S$
```

are equivalent.

Example 13.27

Write an algorithm to find the average mathematics marks of a class of 30 students.

Solution

```
get MARKS
comment MARKS is a sequence of 30 values
        MARKS ($I$) is the marks obtained by $I$ th student
TOTAL ← 0
for $I$ = 1 to 30
    do TOTAL ← TOTAL + MARKS ($I$)
AVERAGE ← $\frac{TOTAL}{30}$
```

13.12 Stepwise Refinement of Algorithm

Algorithms that we wrote down up till now were small. We could write them down in one stroke. All the steps that constitute any of these algorithms were evident and come to our mind quickly. In other words, this means that the problem for which we wrote a particular algorithm was such that we could grasp the problem in its full scope and perspective; it was intellectually manageable to come up with all details of actions or steps that would lead to its solution.

Whenever anybody is faced to solve a problem, it is generally not possible for him/her to think in terms of all minute details of all actions of steps in one stretch. What one is able to do is to come up with broad major steps of actions that will form a first version of an algorithm. After this, one can think of each of these major steps individually. Each such major step is an independent unit; therefore, it becomes intellectually manageable for thinking about the actions it can consist of. This means that one is in a better position to understand and think about more clearly one major step at a time. So one can come up with simpler, detailed and more precise description of subactions that will together achieve the goal of carrrying out the major step. This is what is known as the refinement of the major step. When each major step is refined we get the second version of the algorithm. Each of the substeps needs further refinements. This process can go on till we get the refinement in which each step is sufficiently clear to the person or machine which will execute the algorithm.

We shall illustrate this technique of *stepwise refinement* by two examples.

Example 13.28

Develop an algorithm to find the smallest of three given numbers.

Solution

Suppose a, b, c are three given numbers and we want to decide which of the three is the smallest. What will be our method to achieve our goal ? Our thinking can be as follows. We can take 2 of the 3 given numbers, say a, b and compare them. If $a > b$, a cannot be the smallest of the three and hence the smallest will have to come from b and c. Similarly, if $a < b$, the smallest will come from a and c. So now let us write down this thinking as our first version of the algorithm.

```
get $a, b, c$
if $a > b$
   then smallest will be from $b$ and $c$
   else smallest will be from $a$ and $c$
```

. . (13.10)

At this stage we do not bother to think as to whether b will be smallest or c, when $a > b$. As soon as we know $a > b$ our conclusion is that a is not the smallest and it will come from the remaining two. We thus think step by step.

Now the step (coming after **then**):

```
smallest will be from $b$ and $c$
```

can be made more precise (i.e. refined) as

```
if $b > c$
    then smallest ← $c$
    else smallest ← $b$
```

. . . (13.11)

Similarly, the step (coming after **else**):

smallest will be from a and c

will be refined as

```
if $a > c$
    then smallest ← $c$
    else smallest ← $a$
```

. . . (13.12)

Now incorporating the refinements (13.11) and (13.12) in the first version (13.10) of our algorithm, we get the second version of the algorithm for our problem as given in Fig. 13.17.

```
1. get $a, b, c$
2. if $a > b$
       then if $b > c$
                then smallest ← $c$
                else smallest ← $b$
       else if $a > c$
                then smallest ← $c$
                else smallest ← $a$
3. output smallest
```

Final refined version of algorithm for Example 13.28

Fig. 13.17

The refined version of the algorithm, in Fig. 13.17 is sufficiently detailed and clear that it needs no further refinement and we take it as the final refined version. In step 1, algorithm gets the values of given data, viz. three numbers a, b, c. This is input of the algorithm. In step 2 processing is completed. Note that there are 2 inner **if-then-else** constructs given by (13.11) and (13.12) which are appearing inside the outer **if-then-else** construct that was given by (13.10). This is known as nested appearance of selection – selection inside selection; here 2 selections nested inside 1 selection.

This example was simple and small. Some of you may be able to come up with the final version of the algorithm in one stroke itself. Illustrative examples are chosen to explain the concepts and ideas. It is better that ideas and concepts are explained with small and simple examples. This illustration is one such. Whether problem is simple or complicated, small or big, if you follow the philosophy illustrated by this example, your thinking will remain straightforward and clear; it will not get involved and muddled. You are advised to follow this technique of stepwise refinement. It is described by calling it *top-down design, or divide and conquer strategy.*

Example 13.29

Develop an algorithm to arrange the given set of n numbers in an ascending order.

Solution

Suppose the given numbers are $A(1)$, $A(2)$, . . ., $A(n)$. In this problem, we can think of the procedure as follows. We first find maximum of all given n numbers. That is our first step. Let that maximum occur at the pth element of given sequence. Then we can interchange this pth member $A(p)$ with last member $A(n)$. This is our second step. We want to arrange the given numbers in the ascending order and we have a maximum of the sequence in the last place; so we need not disturb that place here afterwards. So now we can take only first $n - 1$ members of the sequence, viz. $A(1), A(2), \ldots, A(n-1)$ and repeat the two steps that we carried out on the entire sequence, viz. find out a maximum of these $(n-1)$ elements and replace its position with $A(n-1)$. This will bring 2nd largest element at $A(n-1)$ as we want it. It is straightforward that we will have to repeat the 2 steps $(n-1)$ times on shorter and shorter set of elements; each time the length of the sequence we will handle with the 2 steps, will reduce by one as we are building up our final answer from right of the sequence to the left i.e. starting from $A(n)$ and proceeding towards $A(1)$.

Let us write down this thinking as the first version of our algorithm. Notice that we have not intellectually complicated the procedure by not thinking at this stage as to how to find maximum of given elements and how to exchange the positions of two elements. These are the details of the two steps included in the 1st version. Finding a maximum of given sequence of elements and exchanging the positions of 2 elements of a sequence are two broad major steps. The tasks of achieving these will be concentrated during the refinement of these individual steps.

So the first version of the algorithm is

```
get the sequence $A(1), A(2), \ldots, A(n)$
for $k = n$ to 2 by $-1$
    do find max. $A(p)$ from $A(1), A(2), \ldots, A(k)$
       interchange $A(p)$ and $A(k)$
```

$\cdots$ (13.13).

Note in the **for**-construct in (13.13) variable k first takes the value n. Therefore, during 1st execution of this loop max. $A(p)$ will be found from $A(1), \ldots, A(n)$ and will exchange its position with $A(n-1)$ and so on. This is what we wanted and have incorporated it in the algorithm using **for**-construct starting its variable with n and bring it down to 2 using increment of -1 at each step.

Now let us refine the step

find max. $A(p)$ from $A(1), \ldots, A(k)$

that has appeared in (13.13). This means we have to find max. of $A(1), \ldots, A(k)$; locate its position in the sequence and denote that position by p. For achieving this let us now think how we shall locate the max. element from $A(1), \ldots, A(k)$. One way would be to consider a variable with name MAX, assign it value of A (1). This means p, indicating position of maximum, will have value 1. Then compare this value of MAX successively with A (2), A (3),. . ., A (k) . If at any stage any of them is found greater than the value of MAX, assign its value to MAX and change the value of p appropriately. So when we will be through with A (k), appropriate p and correct value of MAX would have been found. Thus, the required refinement is given by

```
$p \leftarrow 1$
MAX $\leftarrow A(p)$
for $J = 2$ to $k$
    do if $(A(J) > \text{MAX})$                           . . . (13.14)
        then $p \leftarrow J$
            MAX $\leftarrow A(p)$
comment $p$th element i.e. $A(p)$ is maximum
        among $A(1), \ldots, A(k)$
```

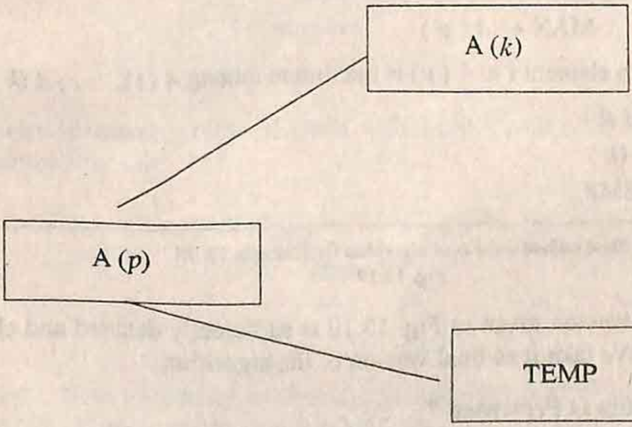


Fig. 13. 18

Now let us refine the step

interchange A (p) and A (k)

that has appeared in (13.13). When working with a computer each of A (p) and A (k) will be in distinct memory locations. If A (k) is moved first at A (p)'s location, A (p) will be lost and will not be available to bring to A (k)'s location. Hence, we use a temporary variable say TEMP, for taking A (p) to its location in first move; then in the second move A (k) is taken to A (p)'s original location; and finally what is in TEMP's location (which is old A (p)) is moved to $A(k)$'s location (See Fig. 13.18). Thus, the required refinement is given by

```
TEMP ← A(p)
A(p) ← A(k)
A(k) ← TEMP
```

... (13.15)

Now incorporating the refinements (13.14) and (13.15), we get the second version of the algorithm as given in Fig.13.19.

```
get the sequence A(1), A(2), . . ., A(n)
for k = n to 2 by 1
   do p ← 1
MAX ← A(p)
for J = 2 to k
      do if (A(J) > MAX)
            then p ← J
                 MAX ← A(p)
 comment pth element i.e. A(p) is maximum among A(1), . . . , A(k)
 TEMP ← A(p)
 A(p) ← A(k)
 A(k) ← TEMP
```

Final refined version of algorithm for Example 13. 29

Fig. 13.19

The second refined version given in Fig. 13.19 is sufficiently detailed and clear and can be executed with ease. We take it as final version of the algorithm.

13.13 How an Algorithm is Presented ?

Up till now we have expressed all our algorithms by writing the instructions of the algorithm one below the other. This we have called as writing algorithms using pseudo language. We have made use of (1) meaningful mnemonic variable names or what are called identifiers, (2) ← as assignment or replacement symbol, (3) sequence, selection and repetition (or iteration) as the basic forms of construction of instructions, (4) **if-then-else** and case constructs for selection, (5) **repeat-until**, **while-do** and **for** constructs.

We are trying to develop in you an ability of constructing an algorithm in a very systematic and disciplined way using this much machinery only; by this you will tend to be more efficient and correct in writing readable algorithm.

There are other instructions available to write algorithms. For example, the algorithm of Fig. 13.16 can be written as

```
get $n$
2. $n \leftarrow n + 1$
   if $n$ is prime
      then output $n$
           stop
      else  go to  step 2
```

... (13.16)

In this we have used a different instruction viz. go to. You may write the above algorithm using go to as in (13.17) also.

```
   get $n$
2. $n \leftarrow n + 1$
   if $n$ is prime
      then go to step 5
      else go to step 2
5. output $n$
```

... (13.17)

If-then-else is one construct. It starts with word **if** and ends after the portion of algorithm that comes after **else**. In

```
if $B$
   then $S$
   else $T$
```

it ends after T. Now no portion of S and T should take control outside this construct. In (13.16). and (13.17) **go to** instruction is taking the control to step 2 or 5 which is outside the construct. When algorithms are small, such things can be done. In large algorithms, if such movements outside the constructs are done, they may endanger correctness and validity of algorithms. Later you may learn mathematical techniques of proving the correctness of algorithms. There these requirements will be mentioned. Good algorithm-developing habits require that you should be disciplined from the beginning and absorb, assimilate and practice good habits from beginning. So we ask you to use constructs given here and do not try to use other constructs like 'go to'. They are not necessary. Their use may turn out to be counter-productive, dangerous and deprive you of good habits and methods of algorithm development.

In Class XI, you have learnt drawing flowcharts to present algorithms. In the beginning they are alright. Again for bigger problems, flowcharts may not inculcate good and sound habits of algorithm development. So we advise you to write your algorithms using pseudo language and writing each construct one below the other, though for some problems in the next section we may give flowcharts, in case you need to refer to them. But concentrate more on the writing of algorithms using pseudo language.

13.14 Some More Examples

Example 13.30

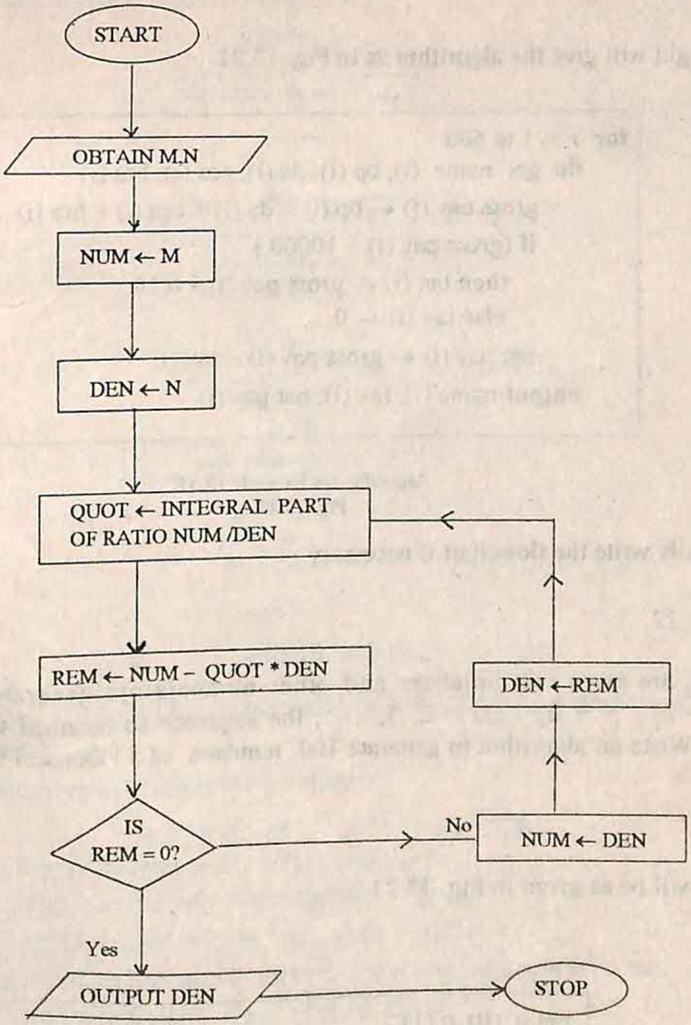
Write an algorithm to obtain HCF of two given positive integers using Euclid's algorithm.

Solution

We know Euclid's algorithm. Therefore we can express it in pseudo language as given in Fig. 13.20. The corresponding flowchart will be as given in Fig. 13.21 (page 681).

```
get $M, N$
comment $M, N$ are two given positive integers
NUM $\leftarrow M$
DEN $\leftarrow N$
QUOT $\leftarrow$ integral part of NUM/ DEN
REM $\leftarrow$ NUM $-$ QUOT $*$ DEN
while REM $\neq 0$
      do NUM $\leftarrow$ DEN
         DEN $\leftarrow$ REM
         QUOT $\leftarrow$ integral part of NUM / DEN
         REM $\leftarrow$ NUM $-$ QUOT $*$ DEN
output DEN
comment HCF is the value of denominator when remainder is zero
```

Algorithm to find HCF of 2 given positive integers
Fig. 13.20



Flowchart for finding HCF of 2 given positive integers
Fig. 13.21

Example 13.31

Name of each person, his basic pay, dearness allowance, house rent allowance and city compensatory allowance are given. Let us assume that if his total salary exceeds Rs 10000, 10% of it is deducted as tax. In all there are 500 persons. Write an algorithm to calculate tax and net pay of each of the person.

Solution

A little thought will give the algorithm as in Fig. 13.22

```
for i = 1 to 500
    do get name (i), bp (i), da (i), cca (i), hra (i)
       gross pay (i) ← bp (i) + da (i) + cca (i) + hra (i)
       if (gross pay (i) > 10000 )
          then tax (i) ← gross pay (i) * 0.10
          else tax (i) ← 0
       net pay (i) ← gross pay (i) – tax (i)
    output name (i), tax (i), net pay (i)
```

Algorithm for Example 13.31

Fig. 13.22

You can easily write the flowchart if necessary.

Example 13.32

If a_0 and a_1 are given two numbers and other numbers are generated using the formula $a_i = a_{i-1} + a_{i-2}$, $i = 2, 3, \ldots$, the sequence so obtained is called *Fibonacci sequence*. Write an algorithm to generate 100 numbers of a Fibonacci sequence using this rule.

Solution

Algorithm will be as given in Fig. 13.23.

```
get a (0), a (1)
for i = 2 to 100
    do a ( i ) ← a ( i – 1) + a (i – 2)
output a
comment output a means all elements of sequence a
```

Algorithm for generating 100 numbers of a Fibonacci sequence

Fig. 13.23

The corresponding flowchart is given in Fig. 13.24.

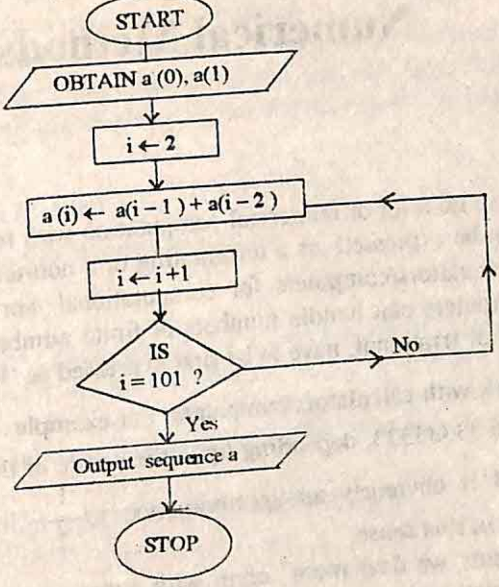


Fig. 13.24

EXERCISE 13.3

1. Use algorithm of Fig. 13.8 as basis and use it to develop an algorithm to find first N primes, given N.
2. Write an algorithm to evaluate the polynomial
$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$
given x and the coefficients of the polynomial.
3. Write an algorithm to multiply two matrices.
4. Write an algorithm to find ${}^nP_r$ and ${}^nC_r$ given n and r.
5. Write an algorithm to find mean and variance of a given set of numbers.
6. Write an algorithm to find the value of $\sin x$ using
$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$
given x and required accuracy of the answer.
7. Write an algorithm to obtain all values of ${}^nC_r$ for $n = 0, 1, \ldots, 10$ using the relation
$${}^nC_r = {}^{n-1}C_r + {}^{n-1}C_{r-1}$$
8. Write separate algorithms to find the sum of specified number of terms of a given arithmetic, geometric and harmonic progression.
9. Develop an algorithm to solve a set of four simultaneous algebraic linear equations.

CHAPTER 14

Numerical Methods

14.1 Introduction

In solving problems we do a lot of numerical computation with real numbers. As you know, every real number can be expressed as a terminating or a non-terminating decimal. We are increasingly using calculators/computers for computational work. As you learnt in the previous chapter, computers can handle numbers of finite number of digits only. Thus, all real numbers, rational or irrational, have to be first expressed as 'finite decimals' for carrying out computational work with calculators/computers. For example $\frac{1}{3}$ has to be expressed as 0.333 or 0.333333 or 0.33333333, depending upon the nature of the problem. When we take 0.333 for $\frac{1}{3}$, 0.333 is obviously an approximation to $\frac{1}{3}$. We refer to 0.333 as an 'approximate number' in this sense.

In practical problems we deal more often with approximate numbers. As a matter of fact any physical measurement may not give the exact value and hence it is an approximate value. When we solve a real problem involving numerical computation, we first decide about the maximum permissible error in the final answer and then determine by certain methods, to which extent the numbers involved, the intermediate computations, etc. are to be approximated. As a thumb rule, if we want the final answer to n places of decimal, we take the numbers involved and the intermediate computational results to $(n + 2)$ places of decimal.

For approximating a number there are two ways. One is to *chop off* (leave out) the extra digits that are not required and the other is to *round off* the last digit to be retained. You are familiar with the rules for rounding off. For example, $\frac{2}{3}$ is 0.6666 if the extra digits are chopped off and is 0.6667 if the fourth digit after the decimal point is rounded off. In both cases $\frac{2}{3}$ is approximated to 4 places of decimal.

Particularly, in problems related to science and technology, we come across various types of equations, integrals, etc. for which we do not have (standard) methods for solution, as we have, for instance, for solving the quadratic equation $ax^2 + bx + c = 0$, $a \neq 0$. In such cases, we try to find their approximate solutions through a series of numerical computations, using certain techniques.

The subject of numerical analysis deals with finding approximate solutions of problems through numerical computation. As you know, computers can perform numerical computations (more often approximate results) very fast, millions per second. The advent of computers thus made numerical methods an important area of study.

The computational work in this chapter is done with a pocket calculator and numerical values are upto 5 places of decimal, extra digits chopped off. Logarithmic tables may also be used for computational work and in that case we should expect the final answer to one or two places of decimal only.

14.2 Successive Bisection Method

This method is based on the following theorem.

Theorem 14.1

If $f(x)$ is continuous in the closed interval $[a, b]$ and $f(a)$, $f(b)$ are of opposite signs, then there is at least one number α in the open interval (a, b) such that $f(\alpha) = 0$.

According to this theorem, if $f(x) = 0$ is a polynomial equation and if we have a and b where $f(a)$ and $f(b)$ are of opposite signs, then there is at least one root of the equation in (a, b).

The successive bisection method is explained below.

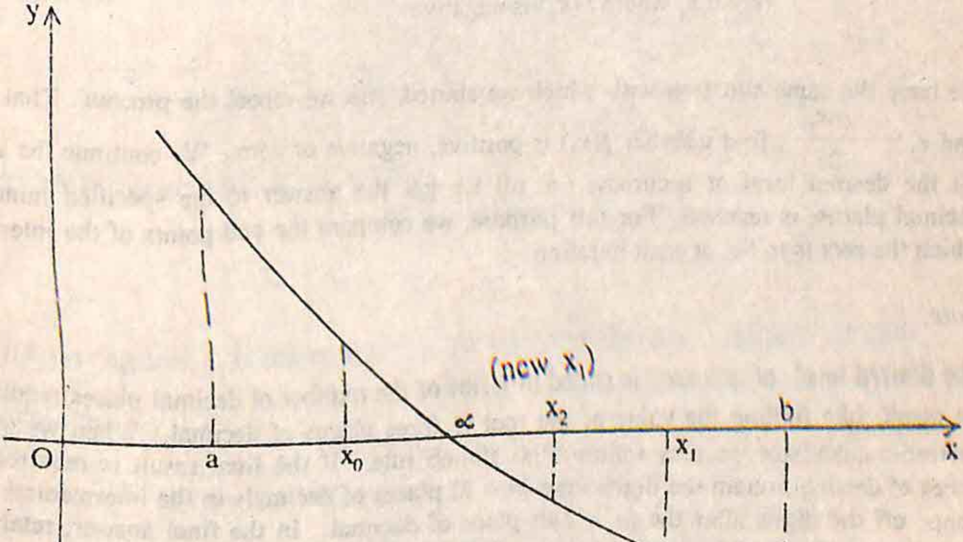


Fig. 14.1

Let the graph of $y = f(x)$ be as shown in Fig. 14.1. Let $f(x_0)$ be positive and let $f(x_1)$ be negative as in Fig. 14.1. According to Theorem 14.1 a root of the equation $f(x) = 0$ lies between x_0 and x_1. The procedure of the successive bisection method is as follows:

Initialisation

We start with x_0 and x_1 where $f(x_0)$ and $f(x_1)$ are of opposite signs.

First iteration

Let
$$x_2 = \frac{x_0+x_1}{2}$$

$f(x_2)$ may be positive, negative or zero. If $f(x_2) = 0$, then x_2 (which is obviously between x_0 and x_1) is the required root. If $f(x_2)$ is positive, then, according to the theorem stated above, the root lies between x_1 (where $f(x)$ is negative) and x_2 (where $f(x)$ is positive). If $f(x_2)$ is negative, then the root lies between x_2 (where $f(x)$ is negative) and x_0 (where $f(x)$ is positive).

Let us assume that $f(x_2)$ is negative. Let us call this x_2 as new x_1.

Second iteration

We have

x_0 where $f(x)$ is positive
(new) x_1 where $f(x_1)$ is negative

We have the same situation with which we started. So we repeat the process. That is, we find $x_2 = \frac{x_0+x_1}{2}$; find whether $f(x_2)$ is positive, negative or zero. We continue the process till the desired level of accuracy, i.e. till we get the answer to the specified number of decimal places, is reached. For this purpose, we compare the end points of the interval, in which the root is to lie, at each iteration.

Note

The desired level of accuracy is stated in terms of the number of decimal places required in the result, like finding the value of the root to three places of decimal. When we use the electronic calculator we may follow this thumb rule. If the final result is required to n places of decimal, retain the digits upto $(n + 2)$ places of decimals in the intermediate steps. Chop off the digits after the $(n + 2)$th place of decimal. In the final answer, retain the digits upto n places of decimal and chop off the other digits. The calculations of this chapter are done with a pocket calculator. If pocket calculator is not available, logarithmetic tables may be used. In that case the desired level of accuracy will have to be less.

Example 14.1

Find a root of the equation $x^3 - x - 4 = 0$ between 1 and 2, to three places of decimal by successive bisection method.

Solution

Initialisation

$$x_0 = 1 \qquad x_1 = 2$$

$$f(x_0) = 1^3 - 1 - 4 \qquad f(x_1) = 2^3 - 2 - 4$$

$$= -4 \qquad = 2$$

$\therefore$ *The root lies between 1 and 2.*

First iteration

$$x_0 = 1 \qquad x_1 = 2 \qquad x_2 = \frac{x_0 + x_1}{2} = 1.5$$

$$f(x_0) = -4 \qquad f(x_1) = 2 \qquad f(x_2) = (1.5)^3 - 1.5 - 4$$

$$= -2.125$$

$\therefore$ *The root lies between 1.5 and 2.*

Since $f(x_2)$ is negative, x_2 becomes new x_0 in the next iteration and x_1 remains the same.

Second iteration

$$x_0 = 1.5 \qquad x_1 = 2 \qquad x_2 = \frac{x_0 + x_1}{2} = \frac{1.5 + 2}{2} = 1.75$$

$$f(x_0) = -2.125 \qquad f(x_1) = 2 \qquad f(x_2) = (1.75)^3 - 1.75 - 4$$

$$= -0.39062$$

Since $f(1.75)$ is negative and $f(2)$ is positive, *the root lies between 1.75 and 2.*

Since $f(x_2)$ is negative, x_2 becomes new x_0 in the next iteration and x_1 remains the same.

Third iteration

$$x_0 = 1.75 \qquad x_1 = 2$$

$$f(x_0) = -0.39062 \qquad f(x_1) = 2$$

$$x_2 = \frac{x_0 + x_1}{2} = \frac{1.75 + 2}{2} = 1.875$$

$$f(x_2) = (1.875)^3 - 1.875 - 4$$

$$= 0.71679 \text{ (Positive)}$$

So the root lies between 1.75 and 1.875.

Since $f(x_2)$ is positive, x_2 becomes the new x_1 in the next iteration and x_0 remains the same.

Fourth iteration

$$x_0 = 1.75 \qquad x_1 = 1.875$$
$$f(x_0) = -0.39062 \qquad f(x_1) = 0.71679$$
$$x_2 = \frac{x_0 + x_1}{2} = \frac{1.75 + 1.875}{2} = 1.8125$$
$$f(x_2) = (1.8125)^3 - 1.8125 - 4$$
$$= 0.14184$$

So the root lies between 1.75 and 1. 8125.

Since $f(x_2)$ is positive, x_2 becomes the new x_1 in the next iteration and x_0 remains the same.

Fifth iteration

$$x_0 = 1.75 \qquad x_1 = 1.8125$$
$$f(x_0) = -0.39062 \qquad f(x_1) = 0.\ 14184$$
$$x_2 = \frac{x_0 + x_1}{2} = \frac{1.75 + 1.8125}{2} = 1.78125$$
$$f(x_2) = (1.78125)^3 - 1.78125 - 4$$
$$= -0.12960$$

So the root lies between 1.78125 and 1.8125.

Since $f(x_2)$ is negative, x_2 becomes the new x_0 in the next iteration and x_1 remains the same.

Sixth iteration

$$x_0 = 1.78125 \qquad x_1 = 1.8125$$
$$f(x_0) = -0.12960 \qquad f(x_1) = 0.14184$$
$$x_2 = \frac{x_0 + x_1}{2} = \frac{1.78125 + 1.8125}{2} = 1.79687$$
$$f(x_2) = (1.79687)^3 - 1.79687 - 4$$
$$= 0.00477$$

So the root lies between 1.78125 and 1.79687.

Since $f(x_2)$ is positive, x_2 becomes the new x_1 in the next iteration and x_0 remains the same.

Seventh iteration

$$x_0 = 1.78125 \qquad x_1 = 1.79687$$
$$f(x_0) = -0.12960 \qquad f(x_1) = 0.00477$$

$$x_2 = \frac{x_0+x_1}{2} = \frac{1.78125 + 1.79687}{2} = 1.78906$$

$$\begin{aligned} f(x_2) &= (1.78906)^3 - 1.78906 - 4 \\ &= -0.06276 \end{aligned}$$

So the root lies between 1.78906 and 1.79687.

Since $f(x_2)$ is negative, x_2 becomes the new x_0 in the next iteration and x_1 remains the same.

Eighth iteration

$$x_0 = 1.78906 \qquad x_1 = 1.79687$$

$$f(x_0) = -0.06276 \qquad f(x_1) = 0.00477$$

$$x_2 = \frac{x_0+x_1}{2} = \frac{1.78906 + 1.79687}{2} = 1.79296$$

$$\begin{aligned} f(x_2) &= (1.79296)^3 - 1.79296 - 4 \\ &= -0.02913 \end{aligned}$$

Therefore, the root lies between 1.79296 and 1. 79687.

Since $f(x_2)$ is negative, x_2 becomes the new x_0 in the next iteration and x_1 remains the same.

Ninth iteration

$$x_0 = 1.79296 \qquad x_1 = 1.79687$$

$$f(x_0) = -0.02913 \qquad f(x_1) = 0.00477$$

$$x_2 = \frac{x_0+x_1}{2} = \frac{1.79296 + 1.79687}{2} = 1.79491$$

$$\begin{aligned} f(x_2) &= (1.79491)^3 - 1.79491 - 4 \\ &= -0.02513 \end{aligned}$$

So the root lies between 1.79491 and 1.79687.

Since $f(x_2)$ is negative, x_2 becomes the new x_0 in the next iteration and x_1 remains the same.

Tenth iteration

$$x_0 = 1.79491 \qquad x_1 = 1.79687$$

$$f(x_0) = -0.02513 \qquad f(x_1) = 0.00477$$

$$x_2 = \frac{x_0+x_1}{2} = \frac{1.79491 + 1.79687}{2} = 1.79589$$

$$f(x_2) = (1.79589)^3 - 1.79589 - 4$$
$$= -\,0.00375$$

So, the root lies between 1.79589 and 1.79687.

Since $f(x_2)$ is negative, x_2 becomes the new x_0 in the next iteration and x_1 remains the same.

Eleventh iteration

$$x_0 = 1.79589 \qquad x_1 = 1.79687$$
$$f(x_0) = -\,0.00375 \qquad f(x_1) = 0.00477$$

$$x_2 = \frac{x_0 + x_1}{2} = \frac{1.79589 + 1.79687}{2} = 1.79638$$

$$f(x_2) = (1.79638)^3 - 1.79638 - 4$$
$$= 0.00050$$

So the root lies between 1.79589 and 1.79638.

Since $f(x_2)$ is positive, x_2 becomes the new x_1 in the next iteration and x_0 remains the same.

Twelfth iteration

$$x_0 = 1.79589 \qquad x_1 = 1.79638$$
$$f(x_0) = -\,0.00375 \qquad f(x_1) = 0.00050$$

$$x_2 = \frac{x_0 + x_1}{2} = \frac{1.79589 + 1.79638}{2} = 1.79613$$

$$f(x_2) = (1.79613)^3 - 1.79613 - 4$$
$$= -0.00167$$

So the root lies between 1.79613 and 1.79638.

Compare the above two limits for the root. The digits in the three places of decimal are the same. Hence the value of the root to three places of decimal is 1.796.

Remarks

(i) How close this value of 1.796 is to the real value of the root can be seen by finding the value of $f(x)$ at this value and comparing it with the theoretical value of 0.

$$f(1.796) = (1.796)^3 - 1.796 - 4$$
$$= -\,0.00279$$

which is very close to 0.

(ii) Had we wanted the root to two places of decimal, we could have stopped at the eighth iteration itself. In that case, the approximate value of the root would be 1.79.

(iii) Had we wanted the root to one place of decimal, we could have stopped at the 6th iteration itself. In that case, the approximaate value of the root would be 1.7

(iv) We started with 1 and 2 at the initialisation stage. Had we started with still closer values, the number of required iterations would have been smaller.

The procedure worked out above can be presented neatly in a tabular form as shown on page 691.

Computation of the Root of $x^3 - x - 4 = 0$ by Successive Bisection Method

Iteration	*Replacement*	x_0	$f(x_0)$	x_1	$f(x_1)$	$x_2 = \frac{x_0 + x_1}{2}$	$f(x_2)$
0		$\underline{1.00000}$	−4.00000	$\underline{2.00000}$	2.00000		
1		1.00000	−4. 00000	$\underline{2.00000}$	2.00000	$\underline{1.50000}$	−2.12500
2	$x_0 \leftarrow x_2$	1.50000	−2.12500	$\underline{2.00000}$	2.00000	$\underline{1.75000}$	−0.39062
3	$x_0 \leftarrow x_2$	$\underline{1.75000}$	−0.39062	2.00000	2.00000	$\underline{1.87500}$	0.71679
4	$x_1 \leftarrow x_2$	$\underline{1.75000}$	−0.39062	1.87500	0.71679	$\underline{1.81250}$	0.14184
5	$x_1 \leftarrow x_2$	1.75000	−0.39062	$\underline{1.81250}$	0.14184	$\underline{1.78125}$	−0.12960
6	$x_0 \leftarrow x_2$	$\underline{1.78125}$	−0.12960	1.81250	0.14184	$\underline{1.79687}$	0.00477
7	$x_1 \leftarrow x_2$	1.78125	−0.12960	$\underline{1.79687}$	0.00477	$\underline{1.78906}$	−0.06276
8	$x_0 \leftarrow x_2$	1.78906	−0.06276	$\underline{1.79687}$	0.00477	$\underline{1.79296}$	−0.02913
9	$x_0 \leftarrow x_2$	1.79296	−0.02913	$\underline{1.79687}$	0.00477	$\underline{1.79491}$	−0.02513
10	$x_0 \leftarrow x_2$	1.79491	−0.02513	$\underline{1.79687}$	0.00477	$\underline{1.79589}$	−0.00375
11	$x_0 \leftarrow x_2$	$\underline{1.79589}$	−0.00375	1.79687	0.00477	$\underline{1.79638}$	0.00050
12	$x_1 \leftarrow x_2$	1.79589	−0.00375	$\underline{1.79638}$	0.00050	$\underline{1.79613}$	−0.00167

Approximate value of the root to three places of decimal is 1.796.

14.3 The Method of False Position

This method is also based on the Theorem 14.1. In the basection method we have taken x_2 to be the mid-point of x_0 and x_1. In the method of false position, x_2 will be the point where the chord through $(x_0, f(x_0))$ and $(x_1, f(x_1))$ intersects the x-axis. The rest of the procedure is similar to the procedure of the bisection method.

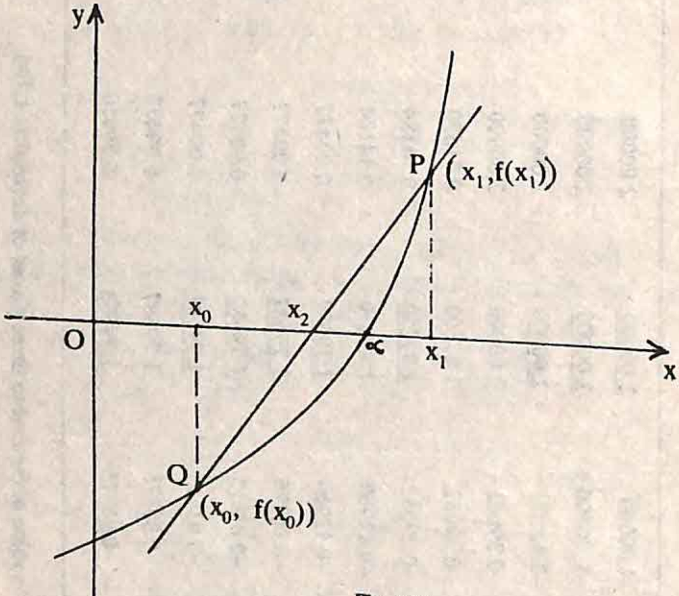


Fig. 14.2

Let $y = f(x)$. Let $f(x_1)$ be positive and $f(x_0)$ negative (Fig. 14.2). According to Theorem 14.1 a root of the equation α lies between x_0 and x_1. Let the line through the points $P(x_1, f(x_1))$ and $Q(x_0, f(x_0))$ intersect the x-axis in x_2.

Equation of the line PQ is

$$y - f(x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_1).$$

This intersects the x-axis in the point given by

$$0 - f(x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_1).$$

$$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)}(x_1 - x_0)$$

If $f(x_2) = 0$, then x_2 is one of the roots. If $f(x_2)$ is positive, then x_2 becomes x_1 (new) in the next iteration. If $f(x_2)$ is negative, then x_2 becomes x_0 (new) in the next iteration. We continue the iterations till the desired level of accuracy for the root is reached, i.e., till we get the answer to the specified number of decimal places. Let us solve the same equation by this method, which we solved by successive bisection method.

Example 14.2

Find a root of the equation $x^3 - x - 4 = 0$ between 1 and 2, to three places of decimal, by the method of false position.

Solution

Initialisation

$$x_0 = 1 \qquad x_1 = 2$$
$$f(x_0) = -4 \qquad f(x_1) = 2$$

$\therefore$ *the root lies between 1 and 2.*

First iteration

$$x_0 = 1 \qquad x_1 = 2$$
$$f(x_0) = -4 \qquad f(x_1) = 2$$

$$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)}(x_1 - x_0)$$

$$= 1 - \frac{-4}{2-(-4)}(2-1)$$

$$= 1.66666$$

$$f(x_2) = (1.66666)^3 - 1.66666 - 4$$

$$= -1.03709$$

Since $f(x_2)$ is negative and $f(x_1)$ is positive, *the root lies between x_2 and x_1 i.e. between 1.66666 and 2.* In the next iteration x_2 becomes the new x_0 because $f(x_2)$ is nagative.

Second iteration

$$x_0 = 1.66666 \qquad x_1 = 2$$
$$f(x_0) = -1.03709 \qquad f(x_1) = 2$$

$$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)}(x_1 - x_0)$$

$$= 1.66666 - \frac{-1.03709}{2-(-1.03709)}(2 - 1.66666)$$

$$= 1.78048$$

$$f(x_2) = (1.78048)^3 - 1.78048 - 4$$
$$= -0.13616$$

So, the root lies between 1.78048 and 2.

Since $f(x_2)$ is negative, in the next iteration x_2 becomes (new) x_0 and x_1 remains the same.

Third iteration

$$x_0 = 1.78048 \qquad x_1 = 2$$
$$f(x_0) = -0.13616 \qquad f(x_1) = 2$$

$$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)}(x_1 - x_0)$$

$$= 1.78048 - \frac{-0.13616}{2 - (-0.13616)}(2 - 1.78048)$$

$$= 1.79447$$

$$f(x_2) = (1.79447)^3 - 1.79447 - 4$$
$$= -0.01606$$

So,the root lies between 1.79447 and 2.

Since $f(x_2)$ is negative, in the next iteration x_2 becomes the (new) x_0 and x_1 remains the same.

Fourth iteration

$$x_0 = 1.79447 \qquad x_1 = 2$$
$$f(x_0) = -0.01606 \qquad f(x_1) = 2$$

$$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)}(x_1 - x_0)$$

$$= 1.79447 - \frac{-0.01606}{2 - (-0.01606)}(2 - 1.79447)$$

$$= 1.79610$$

$$f(x_2) = (1.79610)^3 - 1.79610 - 4$$

$$= -0.00193$$

So, the root lies between 1.79610 and 2.

Since $f(x_2)$ is negative, in the next iteration x_2 becomes (new) x_0 and x_1 remains the same.

Fifth iteration

$$x_0 = 1.79610 \qquad x_1 = 2$$
$$f(x_0) = -0.00193 \qquad f(x_1) = 2$$

$$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)}(x_1 - x_0)$$

$$= 1.79610 - \frac{-0.00193}{2-(-0.00193)}(2 - 1.79610)$$

$$= 1.79629$$

$$f(x_2) = (1.79629)^3 - 1.79629 - 4$$
$$= -0.00028$$

So, the root lies between 1.79629 and 2.

Since $f(x_2)$ is negative, in the next iteration x_2 becomes (new) x_0 and x_1 remains the same.

Sixth iteration

$$x_0 = 1.79629 \qquad x_1 = 2$$
$$f(x_0) = -0.00028 \qquad f(x_1) = 2$$

$$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)}(x_1 - x_0)$$

$$= 1.79629 - \frac{-0.00028}{2-(-0.00028)}(2 - 1.79629)$$

$$= 1.79631$$

$$f(x_2) = (1.79631)^3 - 1.79631 - 4$$
$$= -0.00024$$

So, the root lies between 1.79631 and 2.

Since $f(x_2)$ is negative, in the next iteration, x_2 becomes (new) x_0 and x_1 remains the same.

Seventh iteration

$$x_0 = 1.79631 \qquad x_1 = 2$$
$$f(x_0) = -0.00024 \qquad f(x_1) = 2$$

$$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)} (x_1 - x_0)$$

$$= 1.79631 - \frac{-0.00024}{2 - (-0.00024)} (2 - 1.79631)$$

$$= 1.79633$$

$$f(x_2) = (1.79633)^3 - 1.79633 - 4$$
$$= 0.00007$$

Now $f(x_2)$ is positive and $f(x_0)$ is negative.

Therefore, the root lies between 1.79631 and 1.79633.

The digits in the first three decimal places are the same in the above. Therefore the value of the root upto three places of decimal is 1.796. We got the same answer, which we got under the successive bisection method.

Remark

You can see from the last iteration, that the value of the root to 4 places of decimals is 1.7963.

The procedure worked out above can be nicely represented in a tabular form as given on page 697.

Computation of the Root of $x^3 - x - 4 = 0$ by the Method of False Position

Iteration	*Replacement*	x_0	$f(x_0)$	x_1	$f(x_1)$	$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)}(x_1 - x_0)$	$f(x_2)$
0		1.00000	–4.00000	2.00000	2.00000		
1		1.00000	–4.00000	2.00000	2.00000	1.66666	–1.03709
2	$x_0 \leftarrow x_2$	1.66666	–1.03709	2.00000	2.00000	1.78048	–0.13616
3	$x_0 \leftarrow x_2$	1.78048	–0.13616	2.00000	2.00000	1.79447	–0.01606
4	$x_0 \leftarrow x_2$	1.79447	–0.01606	2.00000	2.00000	1.79610	–0.00193
5	$x_0 \leftarrow x_2$	1.79610	–0.00193	2.00000	2.00000	1.79629	–0.00028
6	$x_0 \leftarrow x_2$	1.79629	–0.00028	2.00000	2.00000	1.79631	–0.00024
7	$x_0 \leftarrow x_2$	1.79631	–0.00024	2.00000	2.00000	1.79633	0.00007

Value of the root to three places of decimal = 1.796

14.4 Newton-Raphson Method

In this method we start with a value x_0 for the root. By some process of trial and error or otherwise we have to choose x_0 such that it is quite close to the real value of the root. Then we consider the tangent to the curve at $(x_0, f(x_0))$ (Fig. 14.3).

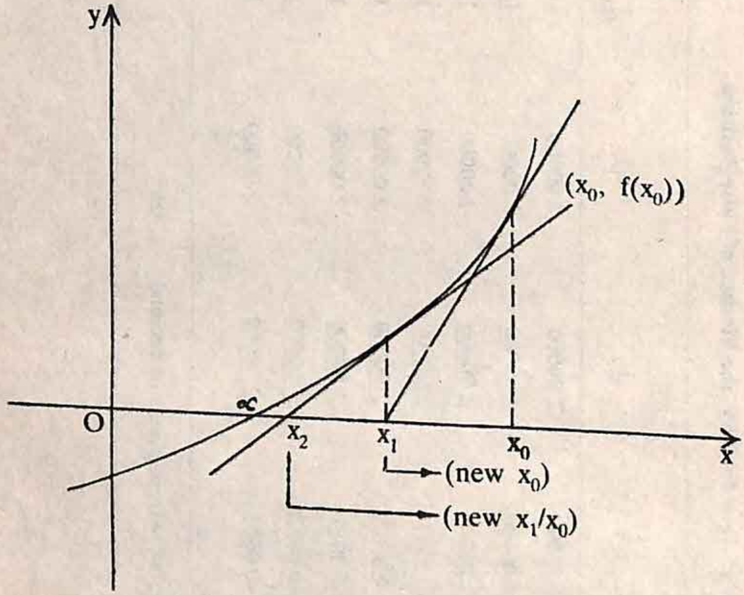


Fig. 14.3

Let this tangent intersect the x-axis in x_1. Now x_1 becomes our new x_0 in the next iteration. Let the tangent at x_1 (new x_0) intersect the x-axis in x_2. Now x_2 is the new x_1 (in turn new x_0) in the next iteration and so on. The process is continued further till we get a value for the root upto the desired level of accuracy, i.e. till we get the answer to the specified number of decimal places.

The equation of the tangent at x_0 is given by

$$y - f(x_0) = f'(x_0)(x - x_0)$$

If this tangent meets the x-axis in x_1, then

$$0 - f(x_0) = f'(x_0)(x_1 - x_0)$$

$$\text{or,} \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

The Newton-Raphson method has the following steps:

1. *Initialisation*

Select x_0 suitably.

2. *First iteration*

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

This x_1 is x_0 for next iteration.

3. *Second iteration*

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

and so on.

The following example illustrates the procedure.

Example 14.3

Find a root of the equation $x^3 - x - 4 = 0$ between 1 and 2 to three places of decimal, by Newton-Raphson method.

Solution

We see that $f(1) = -4$ and $f(2) = 2$. From this we may guess that 2 is nearer to the real root than 1. So let our initial value x_0 be 2.

Initialisation

$$x_0 = 2; \quad f(x) = x^3 - x - 4; \quad f'(x) = 3x^2 - 1$$

First iteration

$$x_0 = 2; \quad f(x_0) = 2; \quad f'(x_0) = 11$$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$= 2 - \frac{2}{11}$$

$$= 1.81819$$

Second iteration

$$x_0 = 1.81819$$

$$f(x_0) = (1.81819)^3 - 1.81819 - 4$$

$$= 0.19240$$

$$f'(x_0) = 3(1.81819)^2 - 1$$

$$= 8.91744$$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$= 1.81819 - \frac{0.19240}{8.91744}$$

$$= 1.79662$$

Third iteration

$$\begin{aligned}
x_0 &= 1.79662\\
f(x_0) &= (1.79662)^3 - 1.79662 - 4\\
&= 0.\ 00258\\
f'(x_0) &= 3(1.79662)^2 - 1\\
&= 8.68018\\
x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)}\\
&= 1.79662 - \frac{0.00258}{8.68018}\\
&= 1.79633
\end{aligned}$$

Comparing the values of x_1 in the third and second iterations, we find that the digits in the first three places of decimal are the same in both. Hence the approximate value of the root to three places of decimal is 1.796.

Remark

The answer that we got here is the same that we got by using successive bisection method and false position method; but the important thing to note is that we could get this answer by fewer iterations in Newton-Raphson method. In using the Newton-Raphson method it is necessary that we should start with an initial value that is close to the real value of the root. In this example choosing 2 as the initial value worked well. The values obtained in successive iterations were showing converging trend. If the selection of the initial value is not done properly, the number of iterations required can be more; even the values in successive iterations may not converge.

The procedure worked out above can be nicely represented in a tabular form below.

Computation of the Root of the Equation $x^3 - x - 4 = 0$ by Newton- Raphson Method

Iteration	x_0	$f(x_0)$	$f'(x_0)$	$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$
Initialisation	2.00000	2.00000	11.00000	
1	2.00000	2.00000	11.00000	1.81819
2	1.81819	0.19240	8.91744	1.79662
3	1.79662	0.00258	8.68018	1.79633

Approximate value of the root to three places of decimal is 1.796

EXERCISE 14.1

1. Find a root of the equation lying in the interval specified against it, using successive bisection method. Give the answer to two places of decimal, chopping off the extra digits.

   (i) $x^3 - 4x + 1 = 0$ (1, 2)

   (ii) $x^3 + x^2 - 1 = 0$ (0.5, 1.0)

   (iii) $x^3 - 3x - 5 = 0$ (2.0, 2.5)

   (iv) $x^3 + x^2 + x + 7 = 0$ (−2.5, −2)

2. Use the method of false position and find the limits within which the root lies at the 5th iteration.

   (i) $x^3 + x^2 - 1 = 0$ (0.5,1.0)

   (ii) $x^3 - 3x - 5 = 0$ (2.0, 2.5)

   (iii) $x^3 + x^2 + x + 7 = 0$ (−2.5, −2)

3. Find a root of the equation lying in the interval specified against it, using Newton-Raphson method. Give the answer to three places of decimal, chopping off the extra digits.

   (i) $x^3 + x^2 - 1 = 0$ (0.5, 1.0)

   (ii) $x^3 - 3x - 5 = 0$ (2.0, 2.5)

   (iii) $x^3 + x^2 + x + 7 = 0$ (−2.5, −2)

   (iv) $x^3 - 4x - 9 = 0$ (2.5, 3)

   (v) $x^3 - 2x - 5 = 0$ (1.75, 2.25)

   (vi) $x^3 - x - 4 = 0$ (1.5, 2.0)

   (vii) $x^3 - 18 = 0$ (2,3)

   (viii) $x^3 - 5x + 3 = 0$ (0.5, 0.75)

   (ix) $x^5 + 5x + 1 = 0$ (− 0.25, − 0.15)

14.5 Solution of a System of Equations

Earlier you learnt how to solve a system of equations with the help of matrices. Here we shall discuss two methods of solving a system of n equations in n variables. We shall assume the system of equations to be consistent and we will find approximate numerical solution to the desired level of accuracy, i.e. till we get the answer to the specified number of decimal places. Let us take a system of three equations and discuss the procedure.

Gauss Elimination Method

Let the given system of equations be

$$a_1x + b_1y + c_1z = d_1 \qquad \ldots (14.1)$$

$$a_2x + b_2y + c_2z = d_2 \qquad \ldots (14.2)$$

$$a_3x + b_3y + c_3z = d_3 \qquad \ldots (14.3)$$

There is no loss of generality, if we assume that at least one of the coefficients a_1 or a_2 or a_3 is not zero, for the reason that if all of them are zero, then the system reduces to a system of 3 equations in 2 variables. Let us take $a_1 \neq 0$ and let also assume that the system has a unique solution.

We will first eliminate x from (14.2) and (14.3). For this purpose we divide (14.1) by a_1 ($\neq 0$, as per our assumption), then multiply it by a_2 and then subtract it from (14.2). You can easily see that x term vanishes. In a similar manner, we eliminate x from (14.3). Thus after carrying out these operations, the given system of equations reduces to the following:

$$a_1x + b_1y + c_1z = d_1 \qquad \ldots (14.1)$$

$$b'_2y + c'_2z = d'_2 \qquad \ldots (14.2)'$$

$$b'_3y + c'_3z = d'_3 \qquad \ldots (14.3)'$$

Here again we may assume, without loss of generality that $b'_2 \neq 0$. Now we will eliminate y from (14.3)′ with the help of (14.2)′. For this purpose we divide (14.2)′ by b'_2 $(\neq 0)$. Then multiply it by b'_3 and then subtract it from (14.3)′. Thus after carrying out this operation, the above system of equations reduces to the following:

$$a_1x + b_1y + c_1z = d_1 \qquad \ldots (14.1)$$

$$b'_2y + c'_2z = d'_2 \qquad \ldots (14.2)'$$

$$c''_3z = d''_3 \qquad \ldots (14.3)''$$

The system of equations is now said to be in the *triangular form*. The process described above is called *triangularisation* of the system.

Now we back-substitute.

From (14.3)″ we have $z = \dfrac{d''_3}{c''_3}$.

Substituting this value of z in (14.2)′, we get the value of y. Substituting the values of y and z in (14.1), we get the value of x. The following example illustrates the procedure.

Example 14.4.

Solve the following system of equations.

$$x + y + z = 3 \qquad \ldots (14.4)$$

$$2x + 3y + z = 6 \qquad \ldots (14.5)$$

$$x - y - z = -3 \qquad \ldots (14.6)$$

Solution

We first eliminate x from (14.5) and (14.6).

$$x + y + z = 3$$

$$(2x + 3y + z) - 2(x + y + z) = 6 - 2 \times 3 \qquad [(14.5) - 2 \times (14.4)]$$

or, $$y - z = 0 \qquad \ldots (14.5)'$$

$$(x - y - z) - (x + y + z) = -3 - 3$$

or, $$-2y - 2z = -6$$

or, $$y + z = 3 \qquad \ldots (14.6)'$$

Thus, the given system is equivalent to the following:

$$x + y + z = 3 \qquad \ldots (14.4)$$

$$y - z = 0 \qquad \ldots (14.5)'$$

$$y + z = 3 \qquad \ldots (14.6)'$$

Now we will eliminate y from (14.6)′ with the help of (14.5)′.

$$y + z - (y - z) = 3 - 0 = 3$$

or, $$2z = 3 \qquad \ldots (14.6)''$$

Thus, the given system reduces to

$$x + y + z = 3 \qquad \ldots (14.4)$$

$$y - z = 0 \qquad \ldots (14.5)'$$

$$2z = 3 \qquad \ldots (14.6)''$$

Now the system is in triangular form. Back-substituting from (14.6)″, we get

$$z = \frac{3}{2} = 1.5$$

Substituting for z in (14.5)′, we get

$$y - 1.5 = 0 \quad \text{or} \quad y = 1.5$$

Substituting for y and z in (14.4), we get

$$x + 1.5 + 1.5 = 3, \quad \text{i.e.} \quad x = 0$$

Thus, the solution is

$$x = 0, y = 1.5, \ z = 1.5$$

Gauss-Seidal Iterative Method

Suppose we have the following system of equations, which is consistent and has a unique solution:

$$a_1 x + b_1 y + c_1 z = d_1 \qquad \ldots (14.7)$$

$$a_2 x + b_2 y + c_2 z = d_2 \qquad \ldots (14.8)$$

$$a_3 x + b_3 y + c_3 z = d_3 \qquad \ldots (14.9)$$

There is no loss of generality if we take $a_1, b_2, c_3 \neq 0$. We can rewrite (14.7) as

$$x = \frac{1}{a_1}(d_1 - b_1 y - c_1 z)$$

i.e. we can write (14.7) as $x = f(y, z)$. Similarly we can write (14.8) as $y = g(z, x)$ and (14.9) as $z = h(x, y)$. Let us write the system of equations in this form.

$$x = f(y, z) \qquad \ldots (14.7)'$$

$$y = g(z, x) \qquad \ldots (14.8)'$$

$$z = h(x, y) \qquad \ldots (14.9)'$$

We can start with some initial solutions. To make calculations simpler we start with the initial solution $y = 0$, $z = 0$. With these initial values of y and z, we can get the approximate value of x from (14.7)′. Let this value be x_1. Now the approximate solution that we have is

$$x = x_1, \quad y = 0, \quad z = 0.$$

Substitute for z and x in (14.8)′, i.e. put $x = x_1$ and $z = 0$. We get an approximate value for y. Let this be y_1. Now the approximate solution that we have is

$$x = x_1, \; y = y_1, \quad z = 0.$$

Substitute $x = x_1$, $y = y_1$ in (14.9)′. We get the approximate value for z. Let this be z_1. Now the approximate solution that we have in the first iteration is

$$x = x_1, \; y = y_1, \quad z = z_1.$$

Now the second iteration begins. Substitute $y = y_1$ and $z = z_1$ in (14.7)′ and get the value x_2 for x. Substitute $x = x_2$ and $z = z_1$ in (14.8)′ and get the value y_2 for y. Substitute $x = x_2$ and $y = y_2$ in (14.9)′ and get the value z_2 for z. Thus at the end of the second iteration we have the solution

$$x = x_2, \; y = y_2, \; z = z_2.$$

Continue the process till you get the solution to the desired level of accuracy. The following example illustrates the procedure.

Example 14.5

Find the solution of the following system of equations to three places of decimal by Gauss-Seidal method.

$$9x + 2y + 4z = 20 \qquad \ldots (14.10)$$

$$x + 10y + 4z = 6 \qquad \ldots (14.11)$$

$$2x - 4y + 10z = -15 \qquad \ldots (14.12)$$

Solution

We rewrite the given system as follows:

$$x = \frac{1}{9}(20 - 2y - 4z) \qquad \ldots (14.10)'$$

$$y = \frac{1}{10}(6 - x - 4z) \qquad \ldots (14.11)'$$

$$z = \frac{1}{10}(-15 - 2x + 4y) \qquad \ldots (14.12)'$$

Initialisation

Let $y = 0$, $z = 0$ in (14.10)′.

First iteration

Putting $y = 0$, $z = 0$ in (14.10)′, we get

$$x = \frac{1}{9}(20) = 2.22222$$

Putting $x = 2.22222$ and $z = 0$ in (14.11)′ we get

$$y = \frac{1}{10}(6 - 2.22222 - 4 \times 0) = 0.377777$$

Putting $x = 2.22222$ and $y = 0.37777$ in (14.12)′, we get

$$z = \frac{1}{10}(-15 - 2 \times 2.22222 + 4 \times 0.37777)$$

$$= -1.79333$$

So at the first iteration, we have

$$x = 2.22222$$
$$y = 0.37777$$
$$z = -1.79333$$

Let us tabulate the solutions obtained in each iteration. See the table at the end of this solution.

Second iteration

From (14.10)′

$$x = \frac{1}{9}[20 - 2 \times 0.37777 + 4 \times 1.79333]$$

$$= 2.93530$$

From (14.11)′

$$y = \frac{1}{10}[6 - 2.93530 + 4 \times 1.79333]$$

$$= 1.02380$$

From (14.12)′ $$z = \frac{1}{10}[-15 - 2 \times 2.93530 + 4 \times 1.02380]$$
$$= -1.68754$$

So, at the end of the second iteration, we have the solution

$$x = 2.93530$$
$$y = 1.02380$$
$$z = -1.68754$$

Third iteration

From (14.10)′ $$x = \frac{1}{9}[20 - 2 \times 1.02380 + 4 \times 1.68754]$$
$$= 2.74472$$

From (14.11)′ $$y = \frac{1}{10}[6 - 2.74472 + 4 \times 1.68754]$$
$$= 1.00054$$

From (14.12)′ $$z = \frac{1}{10}[-15 - 2 \times 2.74472 + 4 \times 1.00054]$$
$$= -1.64873$$

So at the end of third iteration, we have the solution

$$x = 2.74472$$
$$y = 1.00054$$
$$z = -1.64873$$

Fourth iteration

From (14.10)′ $$x = \frac{1}{9}[20 - 2 \times 1.00054 + 4 \times 1.64873]$$
$$= 2.73264$$

From (14.11)′ $$y = \frac{1}{10}[6 - 2.73264 + 4 \times 1.64873]$$
$$= 0.98622$$

From (14.12)′ $$z = \frac{1}{10}[-15 - 2 \times 2.73264 + 4 \times 0.98622]$$
$$= -1.65204$$

So at the end of fourth iteration, we have the solution

$$x = 2.73264$$
$$y = 0.98622$$
$$z = -1.65204$$

Fifth iteration

From (14.10)′ $$x = \frac{1}{9}\,[20 - 2 \times 0.98622 + 4 \times 1.65204]$$
$$= 2.75730$$

From (14.11)′ $$y = \frac{1}{10}\,[6 - 2.75730 + 4 \times 1.65204]$$
$$= 0.98508$$

From (14.12)′ $$z = \frac{1}{10}\,[-15 - 2 \times 2.75730 + 4 \times 0.98508]$$
$$= -1.65742$$

Solution at the end of fifth iteration:

$$x = 2.75730$$
$$y = 0.98508$$
$$z = -1.65742$$

Sixth iteration

From (14.10)′ $$x = \frac{1}{9}\,[20 - 2 \times 0.98508 + 4 \times 1.65742]$$
$$= 2.73994$$

From (14.11)′ $$y = \frac{1}{10}\,[6 - 2.73994 + 4 \times 1.65742]$$
$$= 0.98897$$

From (14.12)′ $$z = \frac{1}{10}\,[-15 - 2 \times 2.73994 + 4 \times 0.98897]$$
$$= -1.65240$$

Solution at the end of sixth iteration:

$$x = 2.73994$$
$$y = 0.98897$$
$$z = -1.65240$$

Seventh iteration

From (14.10)′ $$x = \frac{1}{9}\,[20 - 2 \times 0.98897 + 4 \times 1.65240]$$
$$= 2.73685$$

From (14.11)′ $$y = \frac{1}{10}\,[6 - 2.73685 + 4 \times 1.65240]$$
$$= 0.98727$$

From (14.12)′ $$z = \frac{1}{10}\,[-15 - 2 \times 2.73685 + 4 \times 0.98727]$$
$$= -1.65246$$

Solution at the end of seventh iteration:

$x = 2.73685$
$y = 0.98727$
$z = -1.65246$

Eighth iteration

From (14.10)′ $$x = \frac{1}{9}\,[20 - 2 \times 0.98727 + 4 \times 1.65246]$$
$$= 2.73725$$

From (14.11)′ $$y = \frac{1}{10}\,[6 - 2.73725 + 4 \times 1.65246]$$
$$= 0.98725$$

From (14.12)′ $$z = \frac{1}{10}\,[-15 - 2 \times 2.73725 + 4 \times 0.98725]$$
$$= -1.65255$$

Solution at the end of eighth iteration:

$x = 2.73725$
$y = 0.98725$
$z = -1.65255$

Ninth iteration

From (14.10)′ $$x = \frac{1}{9}\,[20 - 2 \times 0.98725 + 4 \times 1.65255]$$
$$= 2.73730$$

From (14.11)′ $$y = \frac{1}{10}\,[6 - 2.73730 + 4 \times 1.65255]$$
$$= 0.98729$$

From (14.12)′ $$z = \frac{1}{10}\,[-15 - 2 \times 2.73730 + 4 \times 0.98729]$$
$$= -1.65254$$

Solution at the end of ninth iteration:

$$x = 2.73730$$
$$y = 0.98729$$
$$z = -1.65254$$

Compare the values of x, y, z obtained in the eighth and ninth iterations. You find that the digits in the first three places of decimals are the same in both these iterations. Thus, the solution to three places of decimal is

$$x = 2.737$$
$$y = 0.987$$
$$z = -1.652$$

The procedure described above, is presented in the following tabular form:

Solution of the Given System of Equations by Gauss-Seidal Method

Iteration	x	y	z
0	2.22222	0	0
1	2.22222	0.37777	−1.79333
2	2.93530	1.02380	−1.68754
3	2.74472	1.00054	−1.64873
4	2.73264	0.98622	−1.65204
5	2.75730	0.98508	−1.65742
6	2.73994	0.98897	−1.65240
7	2.73685	0.98727	−1.65246
8	$2.\underline{737}25$	$0.\underline{987}25$	$-1.\underline{652}55$
9	$2.\underline{737}30$	$0.\underline{987}29$	$-1.\underline{652}54$

EXERCISE 14.2

1. Solve the following by Gauss elimination method to three places of decimal, chopping off extra digits.

(i)

$$2x + 4y + 2z = 15$$
$$2x + y + 2z = -5$$
$$4x + y - 2z = 0$$

(ii)

$$2x + y + z = 10$$
$$3x + 2y + 3z = 18$$
$$x + 4y + 9z = 16$$

2. Find the solution of the following systems of equations by Gauss-Seidal method, to three places of decimal, chopping off extra digits.

(i)

$$\begin{aligned} x + y &= 2 \\ 3x - 10y &= 3 \end{aligned}$$

(ii)

$$\begin{aligned} 9x + 2y + 4z &= 20 \\ x + 10y + 4z &= 6 \\ 2x - 4y + 10z &= -15 \end{aligned}$$

14.6 Approximations of Some Functions

You are familiar with functions like e^x, $\cos x$, $\log x$, etc. These functions are expressed in the form of infinite power series. For example, you know

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^r}{r!} + \cdots \text{ for all } x$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \text{ for all } x$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \text{ for all } x$$

$$\log(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots \text{ for } |x| \le 1$$

$$\log(1 - x) = -\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots\right) \text{ for } |x| < 1$$

$$\log\left(\frac{1+x}{1-x}\right) = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots\right) \text{ for } |x| < 1$$

We can find approximate values of such function for a given value of x upto a desired level of accuracy, with the help of their power series. There are, of course, other methods also. The following example illustrates the procedure.

Example 14.6

Find the value of e^x for $x = 2$, to three places of decimal.

Solution

We know

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^r}{r!} + \cdots$$

$$i\text{th term} = \frac{x^{i-1}}{(i-1)!}$$

$$(i+1)\text{th term} = \frac{x^i}{i!}$$

$$\frac{(i+1)\text{th term}}{i\text{th term}} = \frac{x^i}{i!} \div \frac{x^{i-1}}{(i-1)!} = \frac{x}{i}$$

$$\therefore \quad (i+1)\text{th term} = \frac{x}{i} \times i\text{th term}$$

We use this formula for finding the value of e^x for a given x, which is 2 in this case.

Table for finding the value of e^x

$$\left(e^2 = 1 + 2 + \frac{2^2}{2!} + \frac{2^3}{3!} + \cdots\right)$$

i	ith term	$(i+1)$th term $= \frac{2}{i} \times i$th term	Sum to $(i+1)$ terms = Sum to i terms + $(i+1)$th term
0	. . .	1.00000	1.00000
1	1.00000	$\frac{2}{1} \times 1.00000 = 2.00000$	3.00000
2	2.00000	$\frac{2}{2} \times 2.00000 = 2.00000$	5.00000
3	2.00000	$\frac{2}{3} \times 2.00000 = 1.33333$	6.33333
4	1.33333	$\frac{2}{4} \times 1.33333 = 0.66666$	6.99999
5	0.66666	$\frac{2}{5} \times 0.66666 = 0.26666$	7.26665
6	0.26666	$\frac{2}{6} \times 0.26666 = 0.08888$	7.35553
7	0.08888	$\frac{2}{7} \times 0.08888 = 0.02539$	7.38092
8	0.02539	$\frac{2}{8} \times 0.02539 = 0.00634$	7.38726

9	0.00634	$\frac{2}{9} \times 0.00634 = 0.00140$	7.38866
10	0.00140	$\frac{2}{10} \times 0.00140 = 0.00028$	7.38894
11	0.00028	$\frac{2}{11} \times 0.00028 = 0.00005$	7.38899
12	0.00005	$\frac{2}{12} \times 0.00005 = .00000$	7.38899

Note: We stop the process here as the contribution of the following terms to the sum upto 3 places of decimal will be zero.

$\therefore$ Value of e^x to three places of decimal

$$= 7.388 \text{ (by chopping off)}$$
$$\text{or } 7.389 \text{ (by rounding off)}$$

EXERCISE 14.3

Find the values of the following, using power series of the function. Give the answer to three places of decimal, chopping off extra digits.

1. $\frac{e^2 + e^{-2}}{2}$
2. e^3
3. sin 0.52
4. cos 0.52
5. sin 0.78
6. cos 0.78
7. $\log_e 5$
8. $\log_e 7$

$\left[\text{Use the expansion of } \log\left(\frac{1+x}{1-x}\right)\right]$

14.7 Numerical Integration

You may recall:

(i) $\int_a^b f(x)\,dx$ represents the area of the region bounded by the curve $y = f(x)$, the x-axis and the ordinates $x = a$ and $x = b$.

(ii) *Fundamental Theorem of Calculus*: If a function $F(x)$ exists such that $F'(x) = f(x)$ in $[a, b]$, then

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

You have evaluated integrals $\int_a^b f(x)\,dx$ by several methods. In all such examples, you were able to find $F(x)$ where $F'(x) = f(x)$. But it is not always so easy and sometimes even not possible to find a $F(x)$ such that $F'(x) = f(x)$. Plenty of such integrals occur in practical problems, particularly in science and engineering. In such situations, integrals are evaluated by numerical methods, to get approximate values to the desired level of accuracy. We will discuss two such methods. These methods are based on the fact that $\int_a^b f(x)\,dx$ is the area of the region bounded by the curve $y = f(x)$, the x-axis and the ordinates $x = a$ and $x = b$.

Trapezoidal Rule

Divide the interval $[a, b]$, into n equal parts or subintervals by means of n points x_0 $(= a)$, x_1, x_2, $\cdots$, x_n $(= b)$.

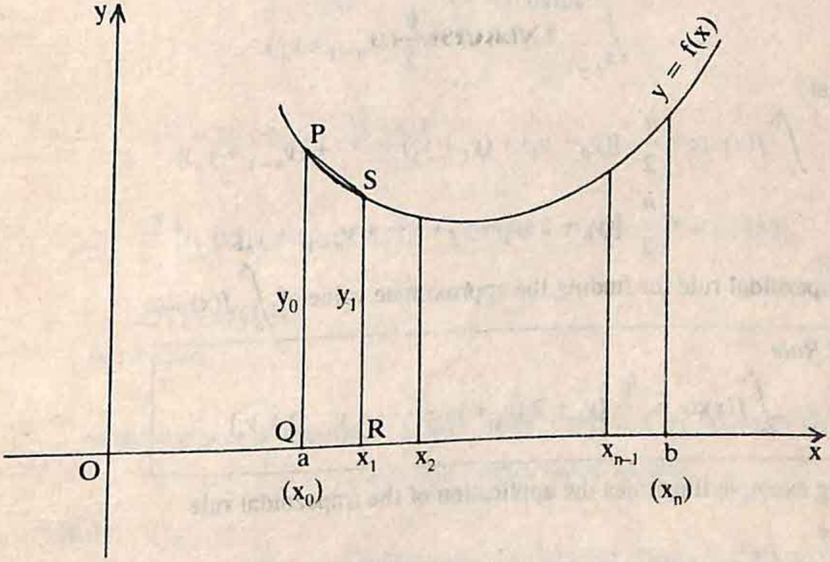


Fig. 14.4

Let the length of each subinterval be h.

So, $$h = \frac{b-a}{n}$$

Now consider the region $PQRS$ bounded by $y = f(x)$, the x-axis, the ordinates $x = x_0$ and $x = x_1$

$$\text{The area of the region} = \int_{x_0}^{x_1} f(x)\,dx.$$

Now consider the trapezium $PQRS$. The area of this trapezoidal region is close to the area of the region $PQRS$ bounded by $y = f(x)$, x-axis and the ordinates $x = x_0$ and $x = x_1$ and we can make it closer still, by taking a suitably large value for n.

If $f(x_0) = y_0$, $f(x_1) = y_1$, then the area of the trapezoidal region $PQRS$ is $\frac{1}{2} h (y_0 + y_1)$.

$$\therefore \quad \int_{x_{0\,(=a)}}^{x_1} f(x)\,dx = \frac{h}{2}(y_0 + y_1).$$

(*Note*: In the above statement the sign '=' stands for 'approximately equal to'.)

Similarly

$$\int_{x_1}^{x_2} f(x)\,dx = \frac{h}{2}(y_1 + y_2)$$

. . . .

. . . .

$$\therefore \quad \int_{x_{n-1}}^{x_{n\,(=b)}} f(x)\,dx = \frac{h}{2}(y_{n-1} + y_n).$$

Adding, we get

$$\int_a^b f(x)\,dx = \frac{h}{2}\left[(y_0 + y_1) + (y_1 + y_2) + \cdots + (y_{n-1} + y_n)\right]$$

$$= \frac{h}{2}\left[(y_0 + 2(y_1 + y_2 + \cdots + y_{n-1}) + y_n\right].$$

This is the trapezoidal rule for finding the approximate value of $\int_a^b f(x)\,dx$.

Trapezoidal Rule

$$\int_a^b f(x)dx = \frac{h}{2}\left[y_0 + 2(y_1 + y_2 + \cdots + y_{n-1}) + y_n\right]$$

The following example illustrates the application of the trapezoidal rule.

Example 14.7

Evaluate by trapezoidal rule $\int_0^2 \frac{dx}{1 + x^4}$, taking $n = 4$. Give the answer to 3 places of decimal.

Solution

$$h = \frac{b - a}{n} = \frac{2 - 0}{4} = 0.5$$

$\therefore$ The points x_0, x_1, x_2, x_3 and x_4 are 0, 0.5, 1.0, 1.5 and 2.0 respectively.

i	x_i	$y_i = \dfrac{1}{1+x_i^4}$
0	0	$\dfrac{1}{1+0} = 1.00000$
1	0.5	$\dfrac{1}{1+(0.5)^4} = 0.94117$
2	1.0	$\dfrac{1}{1+1^4} = 0.50000$
3	1.5	$\dfrac{1}{1+(1.5)^4} = 0.16494$
4	2.0	$\dfrac{1}{1+2^4} = 0.05882$

$$\int_0^2 \frac{dx}{1+x^4} = \frac{h}{2}\,[y_0 + 2(y_1 + y_2 + y_3) + y_4]$$

$$= \left(\frac{0.5}{2}\right)[1.00000 + 2(0.94117 + 0.50000 + 0.16494) + 0.05882]$$

$$= (0.25)\,[4.27104]$$

$$= 1.06776$$

$\displaystyle\int_0^2 \frac{dx}{1+x^4}$ to three places of decimal = 1.067 (after chopping off the extra digits)

= 1.068 (after rounding off)

Simpson's Rule

Let us refer to Fig. 14.4 and the trapezoidal rule. In the interval $[x_0, x_1]$ we considered $f(x)$ to be close to the line through the points $(x_0, f(x_0))$ and $(x_1, f(x_1))$, so that the area of the trapezoidal region *PQRS* is taken to be the approximate value of $\displaystyle\int_{x_0}^{x_1} f(x)dx$.

In Simpson's rule, we consider another type of approximation to $f(x)$ which is explained below.

As in the case of trapezoidal rule divide $[a, b]$ into n sub-intervals of equal length h by means of points $x_0\ (=a), x_1, x_2, \ldots, x_n\ (= b)$.

So, $$x_1 - x_0 = x_2 - x_1 = x_3 - x_2 = \cdots = h.$$

In this method, we will consider a pair of consecutive sub-intervals at a time. So n has to be even.

In Fig. 14.5 only the part of the graph of $f(x)$ for the interval $[x_0, x_2]$ is shown. If P, Q, R are not collinear, there exists a unique parabola, passing through P, Q, R and having its axis vertical. We take this parabola in the interval $[x_0, x_2]$ as a closer approximation to $f(x)$ in that interval.

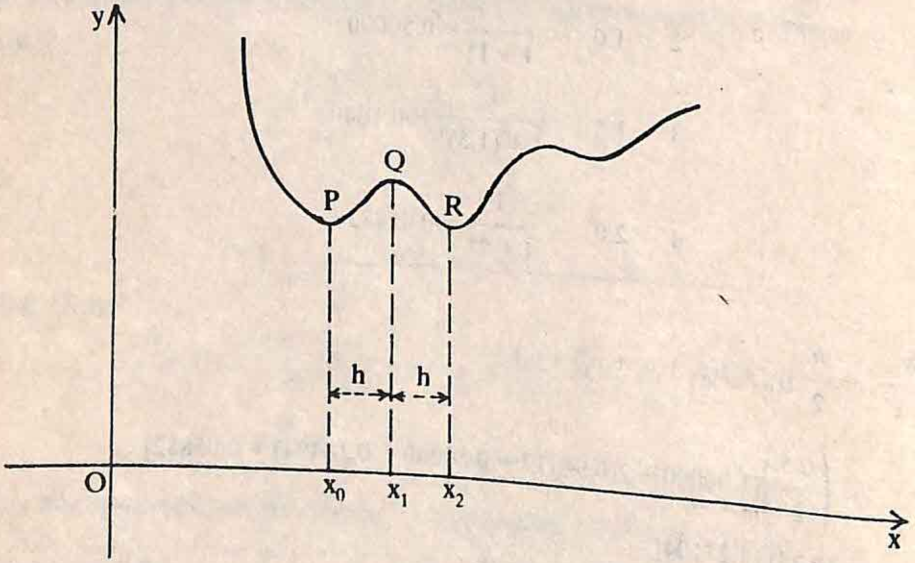


Fig. 14.5

Now the equation of the parabola through P, Q, R can be taken as

$$y = a + b\,(x - x_1) + c\,(x - x_1)^2, \qquad \ldots (14.13)$$

where a, b, c are constants (to be determined).

Since (x_0, y_0), (x_1, y_1), (x_2, y_2) are points on (14.13) we have

$$y_0 = a + b(x_0 - x_1) + c\,(x_0 - x_1)^2 = a - bh + ch^2 \qquad \ldots (14.14)$$

$$y_1 = a + b(x_1 - x_1) + c\,(x_1 - x_1)^2 = a \qquad \ldots (14.15)$$

$$y_2 = a + b(x_2 - x_1) + c\,(x_2 - x_1)^2 = a + bh + ch^2 \qquad \ldots (14.16)$$

(Note: $x_1 - x_0 = x_2 - x_1 = h$)

From (14.15)

$$a = y_1 \qquad \ldots (14.17)$$

From (14.14) and (14.16)

$$y_0 + y_2 = 2a + 2ch^2 = 2y_1 + 2ch^2$$

or,

$$c = \frac{y_0 - 2y_1 + y_2}{2h^2} \qquad \ldots (14.18)$$

From (14.14) and (14.16)

$$y_2 - y_0 = 2bh$$

or,

$$b = \frac{y_2 - y_0}{2h} \qquad \ldots (14.19)$$

Area under the parabola, x-axis and ordinates $x = x_0$ and $x = x_2$

$$= \int_{x_0}^{x_2} [a + b\,(x - x_1) + c\,(x - x_1)^2]\, dx$$

$$= \left[ax + \frac{b\,(x - x_1)^2}{2} + \frac{c\,(x - x_1)^3}{3}\right]_{x_0}^{x_2}$$

$$= a\,(x_2 - x_0) + \frac{b}{2}[(x_2 - x_1)^2 - (x_0 - x_1)^2] + \frac{c}{3}[(x_2 - x_1)^3 - (x_0 - x_1)^3]$$

$$= a \cdot 2h + \frac{b}{2}\,(h^2 - h^2) + \frac{c}{3}\,[h^3 - (-h)^3]$$

$$= 2ah + \frac{2ch^3}{3}$$

$$= 2hy_1 + \frac{2}{3}\left(\frac{y_0 - 2y_1 + y_2}{2h^2}\right)h^3$$

$$= \left(2hy_1 - \frac{2hy_1}{3}\right) + \frac{h}{3}\,(y_0 + y_2)$$

$$= \frac{4hy_1}{3} + \frac{h}{3}\,(y_0 + y_2)$$

$$= \frac{h}{3}\,[y_0 + 4y_1 + y_2]$$

$$\therefore \int_{x_0 (= a)}^{x_2} f(x)\, dx = \frac{h}{3}\,[y_0 + 4y_1 + y_2]$$

(In the above statement the sign '=' stands for 'approximately equal to'.)

In a similar way, we get

$$\int_{x_2}^{x_4} f(x)\,dx = \frac{h}{3}\,[y_2 + 4y_3 + y_4]$$

$$\int_{x_4}^{x_6} f(x)\,dx = \frac{h}{3}\,[y_4 + 4y_5 + y_6]$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$\int_{x_{n-2}}^{x_{n\,(=b)}} f(x)\,dx = \frac{h}{3}\,[y_{n-2} + 4y_{n-1} + y_n].$$

Adding vertically, we get

$$\begin{aligned}\int_a^b f(x)\,dx &= \frac{h}{3}\,[(y_0 + 4y_1 + y_2) + (y_2 + 4y_3 + y_4) \\ &\quad + (y_4 + 4y_5 + y_6) + \cdots + (y_{n-2} + 4y_{n-1} + y_n)] \\ &= \frac{h}{3}\,[y_0 + 4\,y_1 + 2\,y_2 + 4\,y_3 + 2\,y_4 + 4\,y_5 + \cdots + y_n] \\ &= \frac{h}{3}[y_0 + 4\,(\,y_1 + y_3 + \cdots + y_{n-1}) + 2\,(y_2 + y_4 + \cdots + y_{n-2}) + y_n]\end{aligned}$$

This is Simpson's rule for finding the approximate value of $\int_a^b f(x)\,dx$.

Remark

You notice the following structure in the expression within parentheses:

y_0 and y_n occur with coefficient 1.

The remaining y's with even subscripts occur with coefficient 2 and the y's with odd subscripts occur with coefficient 4.

Simpson's Rule

$$\int_a^b f(x)\,dx = \frac{h}{3}\,[y_0 + 4\,(\,y_1 + y_3 + \cdots + y_{n-1}) + 2\,(y_2 + y_4 + \cdots + y_{n-2}) + y_n]$$

We take the same example that we took for illustrating the trapezoidal rule and illustrate the procedure for applying Simpson's rule.

Example 14.8

Evaluate by Simpson's rule $\int_0^2 \frac{dx}{1+x^4}$, taking $n = 4$. Give the answer to 3 places of decimal.

Solution

$$h = \frac{b-a}{n} = \frac{2-0}{4} = 0.5$$

Therefore, $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1.0$, $x_3 = 1.5$, $x_4 = 2.0$.

i	x_i	$y_i = \frac{1}{1+x_i^4}$
0	0	$\frac{1}{1+0} = 1.00000$
1	0.5	$\frac{1}{1+(0.5)^4} = 0.94117$
2	1.0	$\frac{1}{1+(1.0)^4} = 0.50000$
3	1.5	$\frac{1}{1+(1.5)^4} = 0.16494$
4	2.0	$\frac{1}{1+(2.0)^4} = 0.05882$

$$\int_0^2 \frac{dx}{1+x^4} = \frac{h}{3}\,[y_0 + 4\,(y_1 + y_3) + 2y_2 + y_4]$$

$$= \left(\frac{0.5}{3}\right)[1.00000 + 4\,(0.94117 + 0.16494) + (2 \times 0.50000) + 0.05882]$$

$$= 1.08054$$

$$\therefore \quad \int_0^2 \frac{dx}{1+x^4} \text{ to three places of decimal } = 1.080 \quad \text{(after chopping off)}$$

$$= 1.081 \quad \text{(after rounding off)}$$

EXERCISE 14.4

1. Evaluate the integral $\int_0^2 \frac{dx}{1+x}$. Also, evaluate this integral using

   (a) trapezoidal rule taking $n = 4, 8$.

   (b) Simpson's rule taking $n = 4, 8$.

   In each case find your answer to three places of decimal (extra digits dropped) and find the error.

2. Evaluate the integral $\int_0^1 \frac{dx}{1+x^2}$. Also, evaluate the integral using

   (a) trapezoidal rule taking $n = 2, 4$.

   (b) Simpson's rule taking $n = 2, 4$.

   In each case find your answer to three places of decimal (extra digits dropped) and find the error.

3. Evaluate $\int_0^1 \frac{x^p}{x^3 + 12}$ for $p = 0, 1$ using

   (a) trapezoidal rule

   (b) Simpson's rule

   taking $n = 4$. Give your answer to four places of decimal (extra digits dropped).

ANSWERS

EXERCISE 11.1

1. $\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

2. $\{HH, TH, HT1, HT2, HT3, HT4, HT5, HT6, TT1, TT2, TT3, TT4, TT5, TT6\}$

3. $\{HH1, HH2, HH3, HH4, HH5, HH6, TH1, TH2, TH3, TH4, TH5, TH6, HTH, HTT, TTH, TTT\}$

4. Denoting the balls by R_1, R_2, R_3, R_4 the sample spaces may be described as follows:

   (a) $\{R_1, R_2, R_3, R_4\}$ (b) $\{R_1R_2, R_1R_3, R_1R_4, R_2R_3, R_2R_4, R_3R_4\}$

   (c) $\{R_1R_2R_3, R_1R_2R_4, R_1R_3R_4, R_2R_3R_4\}$ (d) $\{R_1R_2R_3R_4\}$

5. Denoting the red balls by R_1, R_2, R_3, R_4 and black balls by B_1, B_2, B_3, the sample spaces are as follows:

   (a) $\{R_1, R_2, R_3, R_4, B_1, B_2, B_3\}$

   (b) $\{R_1R_2, R_1R_3, R_1R_4, R_2R_3, R_2R_4, R_3R_4, B_1B_2, B_1B_3, B_2B_3, R_1B_1, R_1B_2, R_1B_3, R_2B_1, R_2B_2, R_2B_3, R_3B_1, R_3B_2, R_3B_3, R_4B_1, R_4B_2, R_4B_3\}$

6. Red, Red; Red, Black; Black, Red; Black, Black

7. $\{T, H1, H3, H5, H21, H22, H23, H24, H25, H26, H41, H42, H43, H44, H45, H46, H61, H62, H63, H64, H65, H66\}$

EXERCISE 11.2

1. (a) $\{HR_1, HR_2, HR_3\}$ (b) $\{T1, T2, T3, T4, T5, T6\}$

2. $\{B_1G_1, B_1G_2, B_2G_1, B_2G_2, B_3G_1, B_3G_2, G_1G_2\}$

3. (i) $A = \{H2, H4, H6\}$ (ii) $B = \{H2, H3, H5, T2, T3, T5\}$

   (iii) $C = \{T1, T3, T5\}$ (iv) $D = \{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$

4. (i) $A = \{HH, HT\}$ (ii) $B = \{T2, T4, T6\}$

(iii) $C = \{T1, T2, T3, T4, T5, T6, HT\}$

(iv) $D = \{T1, T3, T5\}$

EXERCISE 11.3

1. (a) (i) $A^c = B$ (ii) $B^c = A$ (iii) E^c = getting the sum of the numbers less than 10

(iv) S (v) $\emptyset$ (vi) $\{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3),$
$(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6)\}$

(vii) $\{(1, 1), (1, 2), (1, 3), (1, 4), (3, 1), (3, 2)\}$

(viii) $\{(4, 6), (6, 4), (6, 5), (6, 6)\}$

(ix) $\{(1, 2), (1, 4), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 2), (3, 4),$
$(3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 2), (5, 4), (5, 6), (6, 1),$
$(6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$

(x) $\{(2, 1), (2, 3), (2, 5), (4, 1), (4, 3), (4, 5), (6, 1), (6, 3), (6, 5)\}$

(b) (i) T (ii) T (iii) T (iv) F (v) T (vi) F (vii) F (viii) T (ix) T (x) F

EXERCISE 11.4

1. $\frac{9}{25}$ **2.** (a) $\frac{1}{2}$ (b) $\frac{1}{13}$ (c) $\frac{1}{26}$ (d) $\frac{7}{13}$

3. (i) $\frac{1}{3}$ (ii) $\frac{5}{12}$ (iii) $\frac{1}{4}$ (iv) $\frac{2}{3}$ (v) $\frac{3}{4}$ (vi) 1 (vii) 0

4. (i) $\frac{1}{2}$ (ii) $\frac{1}{2}$ (iii) $\frac{33}{100}$ (iv) $\frac{1}{5}$

(v) $\frac{3}{50}$ (vi) $\frac{47}{100}$ (vii) $\frac{19}{100}$ (viii) $\frac{3}{10}$

5. (i) $\frac{1}{2}$ (ii) $\frac{1}{2}$ (iii) $\frac{1}{4}$ (iv) $\frac{3}{4}$ (v) $\frac{1}{4}$ (vi) $\frac{3}{4}$

6. (i) F (ii) T (iii) T (iv) F

EXERCISE 11.5

1. (i) $\frac{5}{34}$ (ii) $\frac{7}{102}$ (iii) $\frac{35}{102}$ (iv) $\frac{29}{34}$ (v) $\frac{10}{17}$

2. (i) $\frac{5}{33}$ (ii) $\frac{1}{11}$ (iii) $\frac{1}{22}$ (iv) $\frac{2}{11}$ (v) $\frac{5}{22}$ (vi) $\frac{28}{33}$ (vii) $\frac{21}{22}$

3. (i) $\frac{1}{17}$ (ii) $\frac{13}{102}$ (iii) $\frac{1}{221}$ (iv) $\frac{32}{221}$

4. (i) $\frac{91}{228}$ (ii) $\frac{35}{76}$ (iii) $\frac{137}{228}$ 5. $\frac{4}{7}$

EXERCISE 11.6

1. (i) $\frac{1}{4}, \frac{1}{4}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}$

(ii) $\frac{1}{4}$ (iii) $\frac{1}{4}$ 2. (i) The sample space consists of 52 elements with probability $\frac{1}{130}$ each, and 4 elements with probability $\frac{3}{20}$ each; (ii) $\frac{1}{10}$;

(iii) $\frac{1}{5}$; (iv) $\frac{3}{20}$; (v) $\frac{3}{10}$

EXERCISE 11.7

1. (i) $\frac{1}{2}, \frac{3}{4}, \frac{2}{3}, 1$ 2. $\frac{1}{2}, \frac{1}{2}, \frac{3}{7}, \frac{2}{3}$

EXERCISE 11.8

1. (a) and (b) 2. (a)

EXERCISE 11.9

1. (i) $\dfrac{4}{135}, \dfrac{7}{27}$ 2. (a) $\dfrac{3}{8}$ (b) $\dfrac{1}{8}$

EXERCISE 11.10

1. $\dfrac{25}{52}$ 2. $\dfrac{2}{5}$ 3. $\dfrac{33}{118}, \dfrac{55}{118}, \dfrac{15}{59}$

4. $\dfrac{3}{4}$ 5. $\dfrac{27}{83}$ 6. $\dfrac{1}{52}$

EXERCISE 11.11

1. $X(\{r_1, r_2, r_3\}) = 3, X(\{r_2, b_1, b_3\}) = 1$, etc.
   where r_i $(i = 1, 2, 3)$ denote red balls and b_j $(j = 1, 2, 3, 4)$ denote black balls. There are $7^3 = 343$ elements in the sample space.
2. The sample space has $C(7, 3) = 35$ elements and we will have $X(\{r_1, r_2, r_3\}) = 3, X(\{r_1, b_1, b_2\}) = 1$, and so on.

EXERCISE 11.12

1.

14	15	16	17	18	19	20	21
$\frac{2}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{1}{15}$

2. (a) $\dfrac{1}{5}$ (b) $\dfrac{2}{3}$ (c) $\dfrac{7}{15}$

EXERCISE 11.13

1. (i) $\dfrac{3}{32}$ (ii) $\dfrac{7}{64}$ (iii) $\dfrac{63}{64}$ 2. $\dfrac{25}{216}$ 3. $\left(\dfrac{19}{20}\right)^9 \times \dfrac{29}{20}$

4. $\frac{27}{20} \times \left(\frac{19}{20}\right)^7$ **5.** (i) $\left(\frac{1}{4}\right)^5$ (ii) $90 \times \left(\frac{1}{4}\right)^5$ (iii) $\left(\frac{3}{4}\right)^5$

6. (i) $\left(\frac{19}{20}\right)^5$ (ii) $\frac{6}{5} \times \left(\frac{19}{20}\right)^4$ (iii) $1 - \frac{6}{5} \times \left(\frac{19}{20}\right)^4$ (iv) $1 - \left(\frac{19}{20}\right)^5$

7. $\left(\frac{9}{10}\right)^4$

MISCELLANEOUS EXERCISE ON CHAPTER 11

1. (a) Sample space has 36 elements
(1, 1), (1, 2), . . . , (2, 1), (2, 2), . . . , (5, 3), (5, 4), . . . , (6, 5), (6, 6)

(b) {(1, 1), (1, 3), (1, 5), (2, 2), (2, 4), (2, 6), (3, 1), (3, 3), (3, 5),
(4, 2), (4, 4), (4, 6), (5, 1), (5, 3), (5, 5), (6, 2), (6, 4), (6, 6)},
{(1, 2), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 2), (4, 2), (5, 2), (6, 2)},
$E \cup F$, {(1, 1), (1, 3), (1, 5), (3, 1), (3, 3), (3, 5),
(4, 4), (4, 6), (5, 1), (5, 3), (5, 5), (6, 4), (6, 6)}

(c) $\frac{2}{3}, \frac{1}{3}, \frac{13}{36}, \frac{5}{36}$

2. (a) $(E \cap F \cap G^c) \cup (E \cap F^c \cap G) \cup (E^c \cap F \cap G) \cup (E \cap F \cap G)$

(b) $(E \cap F^c \cap G^c) \cup (E^c \cap F \cap G^c) \cup (E^c \cap F^c \cap G)$

(c) $E \cap F^c \cap G^c$

(d) $(E^c \cap F^c \cap G^c) \cup (E \cap F^c \cap G^c) \cup (E^c \cap F \cap G^c) \cup (E^c \cap F^c \cap G)$
$\cup (E \cap F \cap G^c) \cup (E \cap F^c \cap G) \cup (E^c \cap F \cap G) = (E \cap F \cap G)^c$

(e) $(E \cap F \cap G^c) \cup (E \cap F^c \cap G) \cup (E^c \cap F \cap G)$

4. (a) 1 (b) $\frac{2}{5}$ (c) $\frac{1}{5}$ (d) $\frac{1}{5}$

5. $\frac{C(13, 2) \cdot C(11, 3)}{C(24, 5)} = 0.3$ (approx.)

6. $P(E) = \frac{3}{4}, P(F) = \frac{1}{2}, P(E \cap F) = \frac{3}{8}$

7. If event E implies the event F then the probability that out of the two events only F occurs is $P(F) - P(E)$.

8. Yes in all cases. **10.** $\frac{23}{42}$

12. $P(E \cap F \cap G) = P$ (card drawn is a red king) $= \frac{1}{26}$

$$= \left(\frac{1}{2}\right) \left(\frac{4}{26}\right) \left(\frac{1}{2}\right)$$

13. (a) $\frac{1}{2}$ (b) $\frac{1}{2}$

14. $\begin{pmatrix} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ \frac{1}{36} & \frac{2}{36} & \frac{3}{36} & \frac{4}{36} & \frac{5}{36} & \frac{6}{36} & \frac{5}{36} & \frac{4}{36} & \frac{3}{36} & \frac{2}{36} & \frac{1}{36} \end{pmatrix}$

15. (a) $\frac{12}{25}$ (b) $\frac{3}{25}$ (c) $\frac{11}{25}$ (d) $\frac{2}{25}$

16. $\begin{pmatrix} 0 & 1 & 2 & 3 \\ \left(\frac{4}{7}\right)^3 & 3\left(\frac{3}{7}\right)\left(\frac{4}{7}\right)^2 & 3\left(\frac{3}{7}\right)^2\left(\frac{4}{7}\right) & \left(\frac{3}{7}\right)^3 \\ = 0.186 & = 0.420 & = 0.315 & = 0.079 \end{pmatrix}$

17. (i) $\left(\frac{5}{6}\right)^7$ (ii) $35 \times \left(\frac{1}{6}\right)^7$ (iii) $\left(\frac{1}{6}\right)^5$ (iv) $1 - \left(\frac{1}{6}\right)^7$

18. (i) $\left(\frac{3}{4}\right)^4$ (ii) $\left(\frac{1}{4}\right)^4$ (iii) $54 \times \left(\frac{1}{4}\right)^4$ **19.** $\left(\frac{1}{10}\right)^5$

20. (i) $\left(\frac{2}{5}\right)^6$ (ii) $7 \times \left(\frac{2}{5}\right)^4$ (iii) $20 \times \left(\frac{2}{5}\right)^3 \times \left(\frac{3}{5}\right)^3$ (iv) $1 - \left(\frac{2}{5}\right)^6$

21. $\frac{1}{7}$ **23.** $\frac{2}{3}$

EXERCISE 12.1

1. (a) $23 + 4 = 27$ (b) $27 + 59 + 28 + 194 = 308$

MISCELLANEOUS EXERCISE ON CHAPTER 12

1. (e) 17 (f) 45

2. (a) -0.0001, (b) Scatter diagram consists of two straight lines.

(c) Without the scatter diagram we would conclude that there is no relationship. But the situation is quite different.

3. (a) 0.03 (b) The scatter diagram shows points lying close to a curve which first increases and then decreases.

(b) The relationship between the variables is not disclosed by the correlation coefficient because it is not linear.

4. (a) -0.13 (b) 25.6 (c) 0.95 (d) 36.2

5. (b) $b_{vu} = \frac{c}{a} b_{yx}, \quad b_{uv} = \frac{a}{c} b_{xy}$

8. 4 **9.** 0.78 **10.** 2.1

11. $x + 2y - 5 = 0$ **12.** $0.23x + y = 6.21$

13. -0.3162

14. Regression line of y on x is $y = 1.5x + 0.67$.

Regression line of x on y is $x = 0.64\,y - 0.36$

EXERCISE 13.1

1. (i) 307 (ii) 1755 (iii) 1817 (iv) 21845

2. (i) 10110110 (ii) 11100001 (iii) 1111101000
(iv) 10000000000 (v) 10011101110

3. (i) 1100.001 (ii) 11010.01 (iii) 100111.101
(iv) 1001010.0011

4. (i) 5.625 (ii) 3.375 (iii) 10.6875 (iv) 7.9375

5. (i) 158.25 (ii) 278.078125 (iii) 212.140625
(iv) 157.328125

6. (i) 222.4 (ii) 413.2 (iii) 666.1 (iv) 1264.5

7. (i) 2863.3125 (ii) 2748.125 (iii) 451.8125
(iv) 687.625

8. (i) 1***D***1.8 (ii) 416.4 (iii) ***D***89.2 (iv) 1096.*A*

9. (i) 1272, 2*BA* (ii) 3155, 66***D*** (iii) 14667, 19*B*7
(iv) 24723, 29***D***3

10. (i) 10100111 (ii) 1101010111 (iii) 10000110100
11010011010001

11. (i) 10101111001101 (ii) 1010011111011111
(iii) 1010101111001101 (iv) 11100001101111110010

EXERCISE 13.2

1. (i) 1235.6×10^3 (ii) 1.2356×10^3
(iii) $.12356 \times 10^3$ (iv) $.0012356 \times 10^3$

2. (i) 52.463×10^4 (ii) 52.463×10^{-3}
(iii) 52.463×10^{-2} (iv) 52.463×10^{-5}

3. (i) $.124365 \times 10^7$ (ii) $.124365 \times 10^8$
(iii) $.124365 \times 10^5$ (iv) $.124365 \times 10^3$

4. (i) $.54517\ E07$ (ii) $.13054\ E11$
(iii) $.13480\ E06$ (iv) $.17698\ E-01$

5. (i) .278721 E07 (ii) .640486 E05
(iii) .440038 E05 (iv) .142174 E 02

6. (i) .267948 $E - 03$ (ii) .814160 E05
(iii) .165100 E08 (iv) .639870 $E - 06$

7. (i) .385936 E07 (ii) .577119 E10
(iii) .647707 $E - 04$ (iv) .156032 $E - 06$

8. (i) .7083 E05 (ii) .3183 E01
(iii) .8749 $E - 06$ (iv) .2060 E01

9. (i) .2539 E07 (ii) .1251 E08
(iii) .2871 $E - 03$ (iv) .3373 E01

10. (i) 2540687 (ii) 3033010
(iii) 616004 (iv) 4493385

11. (i) 230857 (ii) 5413013
(iii) 693399 (iv) 4887607

12. (i) 10010110010 (ii) 110000110111
(iii) 1001110101100 (iv) 110010000111010

13. (i) 10011001110 (ii) 1100000001010
(iii) 1110101111010 (iv) 110101011101010

EXERCISE 14.1

1. (i) 1.86 (ii) 0.75 (iii) 2.27 (iv) −2.10

2. (i) (.754, 1) (ii) (2.278, 2.5) (iii) (−2.109, −2.101)

3. (i) 0.754 (ii) 2.270 (iii) −2.104
(iv) 2.706 (v) 2.093 (vi) 1.796
(vii) 2.620 (viii) 0.656 (ix) −0.199

EXERCISE 14.2

1. (i) $x = -3.055, y = 6.666, z = -2.777$

(ii) $x = 7, y = -9, z = 5$

2. (i) $x = 1.769, y = 0.231$
   (ii) $x = 2.46, y = 0.71, z = -0.89$

EXERCISE 14.3

1. 3.801
2. 20.085
3. 0.496
4. 0.868
5. 0.703
6. 0.711
7. 1.613
8. 1.942

EXERCISE 14.4

1. (a) 1.116, 1.135 (b) 1.099, 1.098
2. (a) 0.775, 0.782 (b) 0.783, 0.785
3. (a) 0.0815 ($P = 0$), 0.0402 ($P = 1$)
   (b) 0.0816 ($P = 0$), 0.0403 ($P = 1$)

ERRATA

for

Mathematics Textbook for Class XII : Part I

Page No. and Reference	*For*	*Read*
8, 2nd line from top	$[ka + la_{ij}]$	$[ka_{ij} + la_{ij}]$
8, 13th line from top	Notebook – 130 paise each	Notebook – 120 paise each
9, 5th line from top	$B = [b_{ij}]$	$B = [b_{kl}]$
11, 6th line from top	$1 \le k \le n$	$1 \le j \le n$
11, 7th line from top	(k, s) th element	(j, s) th element
11, 9th line from top	$\sum_{j=1}^{n} \sum_{k=1}^{p} a_{ij} b_{kj} c_{ks}$	$\sum_{j=1}^{n} \sum_{k=1}^{p} a_{ij} b_{jk} c_{ks}$
11, 4th line from bottom	$\sum_{j=1}^{n} a_{ik} b_{jk} + \sum_{j=1}^{n} a_{ij} c_{jk}$	$\sum_{j=1}^{n} a_{ij} b_{jk} + \sum_{j=1}^{n} a_{ij} c_{jk}$
15, Q. 20	$-\tan \frac{\alpha}{2}$ (element a_{21})	$\tan \frac{\alpha}{2}$
16, Q. 22	Let $f(x) = x^2 - 5x + 6$. Find $f(A)$ if	Find $A^2 - 5A + 6I$ if
23, 4th line from top	$\sum_{k=1}^{n} a_{ik} C_{jk}$	$\sum_{k=1}^{n} a_{ik} C_{ik}$

Page No. and Reference	*For*	*Read*
31, 6th line from bottom	$\ldots, x_n = \dfrac{D_3}{D}$	$\ldots, x_n = \dfrac{D_n}{D}$
34, 19th line from top	$[-C_1 + C_2, C_3 - C_1]$	$[C_2 + C_1, C_3 - C_1]$
34, 20th line from top	–3 (element in the 3rd row and 3rd column i.e. a_{33})	3
75, Q. 4	$\lim\limits_{x \to 1} \dfrac{x^2 - 1}{x + 1}$	$\lim\limits_{x \to 1} \dfrac{x^2 - 1}{x - 1}$
76. Q. 9	$\lim\limits_{x \to 0} \dfrac{x^4 - 3x^3 + 2}{x^3 - 5x^2 + 3x + 1}$	$\lim\limits_{x \to 1} \dfrac{x^4 - 3x^3 + 2}{x^3 - 5x^2 + 3x + 1}$
84, 19th line from top	*Ignore the second line of the remark:* "Otherwise, the limit will not be 1"	
86, Q. 10	*Ignore this problem with its answer.*	
89, Q. 8	*Ignore the hint against Q. 8*	
98, 19th line from top	examples 2.25 and 2.26	examples 2.16 and 2.17
102, Ex. 2.19	Let $f(x) = \begin{cases} x \text{ if } x \geq 0 \\ x^2 \text{ if } x < 0 \end{cases}$	Let $f(x) = \begin{cases} x \text{ if } x \geq 1 \\ x^2 \text{ if } x < 1 \end{cases}$
120, 12th line from bottom	$\dfrac{d}{dx}(fg) = f\dfrac{dg}{dx} + g\dfrac{df}{g}$	$\dfrac{d}{dx}(fg) = f\dfrac{dg}{dx} + g\dfrac{df}{dx}$
122, Q. 3	$x^{\frac{1}{3}}\, e$	$x^{-\frac{1}{3}}\, e^x$
133, 15th line from top	$\sin\theta = \dfrac{\tan\theta}{\sqrt{1 + \tan^2\theta}} = \dfrac{1}{\sqrt{1 + x^2}}$	$\sin\theta = \dfrac{\tan\theta}{\sqrt{1 + \tan^2\theta}} = \dfrac{x}{\sqrt{1 + x^2}}$
134, Ex. 3.19	Differentiate $\cos^{-1}(4x^3 - 3x)$.	Differentiate $\cos^{-1}(4x^3 - 3x)$, $(-1 < x < 1)$

Page No. and Reference	*For*	*Read*
135, Qs. 2, 3 and 4	$\frac{d}{dx}$ $(\sec^{-1} x)$	$\frac{d}{dx}$ $(\sec^{-1} x)$ for $x > 1$
135, Q. 5	$\operatorname{cosec}^{-1} x$	$\operatorname{cosec}^{-1} x$ for $x > 1$
135, Q. 6	$\sin^{-1} (\cos x)$	$\sin^{-1} (\cos x)$ for $0 < x < \frac{\pi}{2}$
135, Q. 12	$\frac{d}{dx}\left(\cos^{-1}\sqrt{\frac{1+x}{2}}\right) = \frac{-1}{2\sqrt{1-x^2}}$	$\frac{d}{dx}\left(\cos^{-1}\sqrt{\frac{1+x}{2}}\right) = \frac{-1}{2\sqrt{1-x^2}}$, $(-1 < x < 1)$
142, Ex. 3.27	*Replace the second method of solution and the remark on page 143 by the following:*	

Problems on derivatives of inverse trigonometric functions can be easily solved by making suitable substitutions. However, it is found that results may depend on the choice of the substitutions. This is because they may cover different allowed intervals of the independent variable. This is explained in the case of the given problem. The solution to this problem can be tried either by choosing $x = \cos\theta$ or $x = \sin\theta$. We discuss the two cases separately.

Case (i): Let $x = \cos\theta$. Then

$y = \cos^{-1}(4x^3 - 3x) = \cos^{-1}(4\cos^3\theta - 3\cos\theta) = \cos^{-1}\cos 3\theta = 3\theta = 3\cos^{-1}x$.

As the principal value branch of $\cos^{-1}x$ lies in $[0, \pi]$,

$$0 \le 3\theta \le \pi \Rightarrow 0 \le \theta \le \frac{\pi}{3} \Rightarrow \frac{1}{2} \le x = \cos\theta \le 1$$

$$\therefore \frac{dy}{dx} = \frac{d}{dx}(3\cos^{-1}x) = -\frac{3}{\sqrt{1-x^2}} \text{ when } \frac{1}{2} < x < 1$$

Case (ii): Let $x = \sin\theta$. Then

$$y = \cos^{-1}(4x^3 - 3x) = \cos^{-1}(4\sin^3\theta - 3\sin\theta) = \cos^{-1}(-\sin 3\theta)$$

$$= \cos^{-1}\cos\left(\frac{\pi}{2} + 3\theta\right) = \frac{\pi}{2} + 3\theta = \frac{\pi}{2} + 3\sin^{-1}x$$

Using the same argument as given above, we have $0 \le 3\theta + \frac{\pi}{2} \le \pi \Rightarrow -\frac{\pi}{6} \le \theta \le \frac{\pi}{6}$

$$\Rightarrow -\frac{1}{2} \le x = \sin\theta \le \frac{1}{2}$$

$$\therefore \frac{dy}{dx} = \frac{d}{dx}\left(\frac{\pi}{2} + 3\sin^{-1}x\right)$$

$$= \frac{3}{\sqrt{1-x^2}} \text{ when } -\frac{1}{2} < x < \frac{1}{2}.$$

Remark: It can be shown similarly by putting $x = -\cos\theta$ that

$$\frac{dy}{dx} = -\frac{3}{\sqrt{1-x^2}} \text{ when } -1 < x < -\frac{1}{2}.$$

Page No. and Reference	*For*	*Read*
144, Q. 2	$\sin^{-1}(3x-4x^3)$	$\sin^{-1}(3x-4x^3), \left(-\frac{1}{2}<x<\frac{1}{2}\right)$
144, Q. 3	$\sin^{-1}\left(\frac{2x}{1+x^2}\right)$	$\sin^{-1}\left(\frac{2x}{1+x^2}\right), (-1<x<1)$
144, Q. 4	$\cos^{-1}\left(\frac{1-x^2}{1+x^2}\right)$	$\cos^{-1}\left(\frac{1-x^2}{1+x^2}\right), (x>0)$
144, Q. 5	$\sec^{-1}\frac{1}{1-2x^2}$	$\sec^{-1}\left(\frac{1}{1-2x^2}\right), \left(0<x<\frac{1}{\sqrt{2}}\right)$
145, Q. 7	$\sin^{-1}(2x\sqrt{1-x^2})$	$\sin^{-1}(2x\sqrt{1-x^2}),$ $\left(-\frac{1}{\sqrt{2}}<x<\frac{1}{\sqrt{2}}\right)$
146, Q. 7	x	$4x^{\frac{1}{2}}$
150, Q. 14	$\operatorname{cosec}^{-1}\frac{1+x^2}{2x}$	$\operatorname{cosec}^{-1}\left(\frac{1+x^2}{2x}\right), (0<x<1)$
159, 14th line from top	The enclosed area is increasing at the rate of 80 cm^2/s when $r = 10$ cm.	The enclosed area is increasing at the rate of 80π cm^2/s when $r = 10$ cm.
160, 13th line from top	$\frac{ds}{dt} = \frac{5}{2}$	$\frac{ds}{dt} = \frac{5}{2}$ km/h
161, Q. 7	*Ignore this problem with its answer.*	
168, Q. 10	[1, 2]	(1, 2)
172, Fig. 4.14 (b)	*Ignore this figure and its reference.*	
173, Fig. 4.15 (b)	*Ignore this figure and its reference.*	
182, Q. 2 (xii)	*Ignore this problem with its answer.*	
183, Q. 22	maximum	minimum
187, last line	$f(c) = 0$	$f'(c) = 0$
188, Q. 8 (b)	$y = \cos x - 1$ on $[0, 2]$	$y = \cos x - 1$ on $[0, 2\pi]$

Page No. and Reference	*For*	*Read*
191, Q. 5	$ax^2 + bx^2 + cx + d$	$ax^2 + bx^2 + ex + d$
191, Q. 6	$ax^3 + bx^2 + cx + c$	$ax^3 + ex + e$
197, Q. 5	$\sqrt{0.037}$	$\sqrt{0.0037}$
197, Q. 7	x changes from $\frac{\pi}{2}$ to $\frac{22}{14}$	x changes from $\frac{22}{14}$ to $\frac{\pi}{2}$
220, Q. 3	$\sim(\sim p \wedge \sim q) = p \wedge q$	$\sim(\sim p \wedge \sim q) = p \vee q$
221, 1st line from top	and $\sim(\sim p \vee \sim q) = p \vee q$	and $\sim(\sim p \vee \sim q) = p \wedge q$
221, 14th line from top	$\sim(p \vee q) = \sim p \vee \sim q$	$\sim(p \vee q) = \sim p \wedge \sim q$
240, Q. 1	$p + \{[p' \,.\, (p + q)] + (q \,.\, r)\}$	$p + \{[p' \,.\, (p + q)] + (q \,.\, p)\}$.
243, Q. 15	*Ignore this problem with its answer.*	
245, Q. 12 (i)	$\begin{bmatrix} 2a & 0 \\ 0 & 2a \end{bmatrix}$.	$\begin{bmatrix} 2a & 2b \\ 0 & 2a \end{bmatrix}$
246, Q. 14 (iii)	$[a^2 + b^2 + 0 + d^2 + ac + bd]$	$[a^2 + b^2 + c^2 + d^2 + ac + bd]$.
247, Q. 1 (iii) of Ex. 1.4	$-6, -2, 5$ in the third row	6, 2, 5 in the third row
247, Q. 2 (i) of Ex. 1.4	$\frac{1}{7}\begin{bmatrix} 3 & 1 \\ -1 & 2 \end{bmatrix}$	$\frac{1}{8}\begin{bmatrix} 3 & 2 \\ -1 & 2 \end{bmatrix}$
247, Q. 2 (vi) of Ex. 1.4	$-\begin{bmatrix} -2 & 0 & -1 \\ -9 & -2 & 3 \\ -6 & -1 & 2 \end{bmatrix}$.	$\begin{bmatrix} -2 & 0 & 1 \\ 9 & 2 & -3 \\ 6 & 1 & -2 \end{bmatrix}$
247, Q. 4 of Ex. 1.5	the given answer	$t = -\frac{7}{2}$; no.
248, Q. 8 of Ex. 2.1	$\left[0, \frac{3}{2}\right]$	$\left[0, \frac{2}{3}\right]$
248, Q. 1 of Ex. 2.3	the given answers	x, x^2
248, Q. 2 of Ex. 2.3	the given answers	$\sin(\sin x)$, $4x$, not the same
249, Q. 9 of Ex. 2.6	*Ignore this problem with its answer.*	
250, Q. 11 of Ex. 3.2	$\sec^2 x + 2\cos x - \frac{1}{2x} - e^x$	$\sec^2 x + 2\cos x - 3\sin x - \frac{1}{2x} - e^x$
250, Q. 2 of Ex. 3.4	$\frac{1 + \tan x + x\sec^2 x}{(1 + \tan x)^2}$	$\frac{1 + \tan x - x\sec^2 x}{(1 + \tan x)^2}$

Page No. and Reference	For	Read
251, Q. 4 of Ex. 3.8	$e^x - \cos^3 x \sin^2 x \times$ $[1 - 3\tan x + 2\cot x]$	$e^x \cos^3 x \sin^2 x \times$ $[1 - 3\tan x + 2\cot x]$
252, Q. 8 of Ex. 3.9	$\left(\dfrac{\cos\theta + 2\cos 2\theta}{\sin\theta + 2\sin 2\theta}\right)$	$-\left(\dfrac{\cos\theta + 2\cos 2\theta}{\sin\theta + 2\sin 2\theta}\right)$
252, Q. 6 of Ex. 3.11	$-\dfrac{1 + \log x}{(x + \log x)^2}$	$-\dfrac{1 + \log x}{(x \log x)^2}$
253, Qs. 8 to 22 of Misc. Ex. on Chapter 3	*Ignore Q. 8 with its answer and replace Qs. 9, 10, 11, ..., 22 by 8, 9, 10, ..., 21 respectively.*	
253, Qs. 23 (i) and (ii)	23 (i) $2x \sin 2x^2$ (ii) $\dfrac{1}{2} \sec^2 \dfrac{x}{2}$	22 (i) $2x \sin 2x^2$ (ii) $\dfrac{1}{2} \sec^2 \dfrac{x}{2}$
253, Qs. 25, 29, 31	*Replace Qs. 25, 29 and 31 by 24, 28 and 30 respectively.*	
254, Q. 3 of Ex. 4.1	$a = 3, b = -2, c = 6$	$a = 3, b = -2, c = 5$
254, Q. 1 of Ex. 4.2	12250	12250 m
254, Q. 6 of Ex. 4.3	$-\dfrac{8}{3}$ m/s	$\dfrac{8}{3}$ m/s
254, Q. 7 of Ex. 4.3	*Ignore this answer.*	
254, Q. 3 of Ex. 4.3	$\dfrac{1}{\pi}$ cm^3/s	$\dfrac{1}{\pi}$ cm/s
254, Q. 10 of Ex. 4.4	-4	-2
254, Qs. 1 (i) to (ix)	the given answers	1. (i) Min. = 3 (ii) Max. = 10 (iii) Min. = −2 (iv) No max. or no min. (v) Min. = 0 (vi) Max. = 3 (vii) Min. = 4, Max. = 6 (viii) Max. = 4, Min. = 2 (ix) Max. = sin 1, Min. = −sin 1

Page No. and Reference	*For*	*Read*
255, Q. 2 (xii)	*Ignore this answer.*	
255, Qs. 2 (xiv) to (xix)	*In place of the given answers, read the following:* 2. (xiv) Min. at $x = \frac{\pi}{4}$, Value $= \frac{1}{2}$ (xv) Min. at $x = -\frac{\pi}{6}$, Value $= -\frac{\sqrt{3}}{2} + \frac{\pi}{6}$ Max. at $x = \frac{\pi}{6}$, Value $= \frac{\sqrt{3}}{2} - \frac{\pi}{6}$ (xvi) Min. at $x = 3$, Value $= 0$ (xvii) Min. at $x = 1$, Value $= 0$ Max. at $x = \frac{3}{5}$, Value $= \frac{108}{3125}$ (xviii) Min. at $x = \frac{\pi}{4}$, Value $= -\frac{1}{512}$ (xix) Min. at $x = -1$, Value $= 0$ Max. at $x = -\frac{1}{5}$, Value $= \frac{3456}{3125}$	
255, Q. 5 (iii)	Min. $= -1.77$ Max. $= 19.625$	Min. $= -1.75$ Max. $= 19.625$
255, Qs. 7, 8	7. Min. at $x = 2$, Value $= -39$ Max. at $x = 3$, Value $= 16$ 8. Min. -55.8 Max. $= 321$	7. Min. at $x = 2$, Value $= -39$ Max. at $x = 0$, Value $= 25$ 8. Min. $= -63$ Max. $= 257$
255, Q. 22	The cylinder with radius $\left(\frac{50}{\pi}\right)^{\frac{1}{3}}$	The cylinder with radius $\left(\frac{50}{\pi}\right)^{\frac{1}{3}}$ cm and height $2\left(\frac{50}{\pi}\right)^{\frac{1}{3}}$ cm
256, missing Qs. 5, 6 of Ex. 4.7	*Write the following answers of Qs. 5 and 6:* 5. $c = \frac{1}{2}$, 6. $c = \frac{1}{\sqrt{3}}$	
257, Q. 8	4π	4π cm^2

Page No. and Reference	*For*	*Read*
264, Q. 13	13	14
264, Q. 14	*Ignore Q. 14 (i) and (ii) with circuit diagrams.*	
264, Q. 16	16	17
264, missing Q. 16	*Write the answers of Q. 16 as follows:* 16. (i) $(p+q).(p+r)$ (ii) $(r+q).(r+p).(s+p).(s+q)$	

Answers of Exercise 1.7

(i) $x = \frac{23}{19}, y = -\frac{32}{19}$

(ii) $x = 1,\ y = -\frac{1}{2}$

(iii) $x = -8,\ y = 5$

(iv) $x = \frac{11}{24}, y = \frac{1}{24}$

(v) $x = 1,\ y = 2, z = -1$

(vi) $x = \frac{1}{2}, y = \frac{3}{2}, z = -1$

(vii) $x = 1,\ y = 2, z = 1$

(viii) $x = 2,\ y = 1, z = 3$

$$P\,(H_i\,/\,E\,) = \frac{P\,(H_i)\;P\,(E/H_i)}{\sum P\,(H_i)\;P\,(E/H_i)}$$

1261